Análise da informatividade de microgrupos em Mapas Auto-Organizáveis para identificação de variáveis importantes no diagnóstico de COVID-19

Vinícius G. P. Grande[1], Leandro A. da Silva[2]

[1]Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie – São Paulo, SP - Brazil

[2]Programa de Pós-Graduação em Engenharia Elétrica e Computação – Universidade Presbiteriana Mackenzie – São Paulo, SP - Brazil

vinygp8@gmail.com, leandroaugusto.silva@mackenzie.br

Abstract. *This paper describes a study based on the classification results of a machine learning algorithm in relation to hemograms of patients with suspected COVID-19. Based on the hypothesis that the most important variables in the blood count are those that have less informativeness in Self-Organizing Maps, we sought to identify, through the SOMLI-KNN algorithm, the informativeness metric and perform microgroup analysis of the Self-Organizing Map so that this hypothesis can be validated. As a result, the new approach showed rapid execution and high accuracy in classifying the diagnosis of patients with suspected COVID-19 compared to other classifiers in the literature.*

Resumo. *Este trabalho descreve um estudo feito com base nos resultados da classificação de um algoritmo de aprendizado de máquina em relação a hemogramas de pacientes com suspeita de COVID-19. Com base na hipótese de que as variáveis mais importantes do hemograma sejam aquelas que possuem menos informatividade em Mapas Auto-Organizáveis, buscou-se identificar através do algoritmo SOMLI-KNN a métrica de informatividade e fazer análise em microgrupos do Mapa Auto-Organizável para que essa hipótese possa ser validada. Com isso, a nova abordagem apresentou rápida execução e alta precisão na classificação do diagnóstico de pacientes com suspeita de COVID-19 em comparação com outros classificadores da literatura.*

1. Introdução

A COVID-19 é uma doença causada pelo coronavírus, conhecido como SARS-CoV-2, que apresenta um espectro clínico variando de infecções assintomáticas até quadros graves. A maior parte dos pacientes com COVID-19 tem a possibilidade de possuírem poucos sintomas da doença ou serem assintomáticos, visto que aproximadamente 20% dos casos detectados requerem atendimento hospitalar devido a dificuldades respiratórias [Ministério da Saúde 2019].

A Organização Mundial da Saúde recebeu um alerta em Dezembro de 2019 a respeito de diversos casos de pneumonia que ocorreram na cidade de Wuhan, localizada na China. Após uma semana, as autoridades locais atestaram que um novo tipo de coronavírus havia sido constatado. Em março de 2020 a disseminação da COVID foi caracterizada como uma pandemia, visto que ela atingiu vários países e regiões do mundo [Organização Pan-Americana da Saúde 2020].

Os principais tipos de testes para COVID-19 são o RT-PCR, a sorologia e os exames rápidos de antígeno e anticorpos. A comprovação de COVID-19 é obtida por meio da detecção do RNA do SARS-CoV-2 no paciente analisado. A sorologia analisa a resposta imunológica do corpo em relação ao vírus e os testes rápidos de antígeno e anticorpos, que respectivamente fazem a detecção de proteínas na fase de atividade da infecção e realizam a identificação de uma resposta imunológica do corpo em relação ao vírus [Fleury, 2020].

Porém, estes testes principais para COVID-19 possuem a desvantagem de terem um custo e tempo elevado. Dessa forma, é importante observar que os hemogramas podem ser uma alternativa com custo acessível e rápida para fazer a distinção entre pacientes com COVID-19 positivo e negativo.

Deste modo, outros estudos investigaram dados de hemogramas de pacientes utilizando diferentes abordagens com algoritmos de Machine Learning para identificação de pacientes com COVID-19, tendo como exemplo abordagens utilizando o algoritmo de *Boosting SMOTE* e máquinas de vetor de suporte, além do classificador *Naive Bayes* e outros métodos (redes neurais, florestas aleatórias, regressão logística e máquinas de vetor de suporte)[Cabitza et al. 2020];[Soares et al. 2020]; [de Moraes Batista et al. 2020]; [Avila et al. 2020] [Souza et al. 2021]

Uma das lacunas deixadas pelos trabalhos anteriores decorre do fato de que hemogramas tem diversas variáveis e que para a medicina é importante saber a relevância de cada uma delas no diagnóstico de COVID-19.

A proposta consiste em analisar os microgrupos gerados pelos Mapas Auto-Organizáveis propostos por Kohonen (2013) e utilizá-los em conjunto com o classificador *LI-KNN* definido por Song et al. (2007) para criar um novo algoritmo de classificação intitulado de *SOMLI-KNN,* que auxiliará na predição do diagnóstico de COVID-19.

Através dessa análise, poderá ser observado que os fatores mais importantes para o diagnóstico de COVID-19 correspondem às variáveis com menos informatividade dos microgrupos dos Mapas Auto-Organizáveis.

Além da introdução, o trabalho está organizado em referencial teórico com explicação do coronavírus, do algoritmo *KNN*, do conceito de informatividade e de Mapas Auto-Organizáveis na seção 2. Em seguida é apresentada a metodologia da pesquisa na seção 3 e os resultados e análises na seção 4. Por último, na seção 5, a conclusão e trabalhos futuros são apresentados.

2. Referencial Teórico

2.1. Coronavírus

O Coronavírus causa diversas doenças tanto em seres humanos como animais, especialmente nas vias respiratórias. O coronavírus pertence à família *Coronaviridae* e é um vírus de RNA positivo de cadeia única rodeado por um envelope. Está dividido em quatro géneros: *Alpha-*, *Beta-*, *Gamma-*, e *Deltacoronavirus*. Os infectados com SARS-CoV-2 podem apresentar diversos sintomas que vão desde sintomas leves a graves. Falta de ar, tosse e febre são os sintomas mais comuns relatados em 31, 82, e 83% dos pacientes. [Wang et al. 2020]; [Ciotti et al. 2020].

Com o intuito de facilitar a distinção de pacientes com diagnóstico de COVID-19 positivo e negativo em uma admissão hospitalar, pode-se realizar um diagnóstico laboratorial utilizando RT-PCR, testes rápidos de antígeno e anticorpos, sorologia e hemogramas.

2.2. Algoritmo *KNN*

Na área de classificação de dados, o algoritmo *KNN* ou K Vizinhos mais Próximos (do inglês *K Nearest Neighbors*) é um dos algoritmos considerados como referência na literatura [Han; Kamber; Pei 2012]; [Wu e Kumar 2009]; [Cover e Hart 1967]. Durante o processo de classificação dos dados é necessário fazer um cálculo da distância entre o objeto em análise e cada objeto armazenado no conjunto de treinamento, realizar a ordenação dos objetos armazenados pelas menores distâncias e no final classificar o objeto com a classe que é maioria dos vizinhos mais próximos [Wettschereck 1994]. O amplo uso deste classificador se deve principalmente a sua acurácia elevada em problemas de classificação de dados e facilidade de implementação [Wu e Kumar 2009]; [Maimon e Rokach 2010].

2.3. Informatividade

O conceito de informatividade é que dois objetos são capazes de compartilhar o mesmo rótulo de classe quando sua distância é pequena o suficiente, tendo assumido que estes objetos possuem uma distribuição uniforme. Essa ideia é a mesma presente na classificação do algoritmo *KNN*. Em compensação, comparado com os algoritmos de classificação tradicionais que calculam as distâncias entre pares entre o objeto de consulta e os vizinhos, a métrica baseada em informatividade também mede a proximidade entre os objetos vizinhos, ou seja, os objetos informativos devem ter uma grande distância de objetos diferentes. Isso garante que os locais de outros objetos informativos tenham a mesma probabilidade de classificação da mesma classe. [Song et.al 2007]. Este parâmetro foi analisado com o uso de bases de dados artificiais e bases de dados da literatura através de um estudo comparativo com outros algoritmos: *KNN*, *SVM* e *Random Forest*. Os conjuntos de dados foram gerados artificialmente simulando diferentes tipos de distribuições de objetos em duas classes com o intuito de analisar a sensibilidade deste parâmetro nestes tipos de dados. Sendo assim, foi feita a parametrização deste valor e percebeu-se o seu potencial em situações de sobreposição de classes[Grande e Silva 2019].

2.4. Mapas Auto-Organizáveis

Mapa Auto Organizável (*SOM,* do inglês *Self-Organizing Map*) é um método automático de análise de dados amplamente aplicado a problemas de agrupamento e exploração de dados. *O SOM* está relacionado à quantização vetorial usada no processamento e transmissão de sinais digitais. Como na quantização vetorial , o *SOM* representa uma distribuição de objetos de dados de entrada usando um conjunto finito de modelos. No *SOM*, no entanto, esses modelos são automaticamente associados aos nós de uma grade regular (geralmente bidimensional) de forma ordenada, de modo que objetos mais semelhantes sejam associados automaticamente aos nós adjacentes à grade, enquanto modelos menos semelhantes são situados mais distantes um do outro na grade. Essa organização, uma espécie de diagrama de similaridade dos modelos, permite obter um *insight* sobre as relações topográficas dos dados, especialmente de itens de dados de alta dimensão.

Essa abordagem foi utilizada em conjunto com o algoritmo *LI-KNN* (Locally *Informative K-Nearest Neighbor*) [Song et.al 2007] através de um classificador híbrido denominado *SOMiNN*, que explora os conceitos de quantização, manutenção de topologia e informatividade [Moreira e Silva 2017]; e com o algoritmo *KNN* implementado em duas versões: *SOM-KNN* e *SOM4-KNN* [Silva e Del-Moral-Hernandez 2011].

Sendo assim, pretende-se, neste projeto, medir a informatividade dos objetos nos microgrupos presentes nos Mapas Auto-Organizáveis, verificando quais deles estão mais sobrepostos, ou seja, possuem menos informatividade nos Mapas Auto-Organizáveis [Kohonen 2013].

3. Metodologia

A linguagem de programação R e o ambiente de desenvolvimento RStudio foram utilizados para implementar o algoritmo *SOMLI-KNN*, realizar experimentos de geração de protótipos por SOM e criar os gráficos apresentados neste trabalho.

Os dados de exames médicos de COVID-19 obtidos pelo Instituto Fleury foram usados como conjunto de dados. Essas informações foram padronizadas com a fórmula *z-score* para evitar que os experimentos fossem enviesados.

Após a análise dos resultados obtidos com o *SOMLI-KNN*, o algoritmo foi comparado com outros classificadores da literatura: *KNN* e *LI-KNN*. A parametrização escolhida para estes algoritmos foi com o valor de K igual a oito e o valor de I equivalente a um.

O desempenho do estudo comparativo foi feito com a utilização da metodologia de validação cruzada *k-fold* e com apresentação dos resultados através de uma matriz de confusão. A métrica que foi utilizada neste estudo é a acurácia de acertos do algoritmo.

Pode-se destacar que os valores das dimensões dos Mapas Auto-Organizáveis gerados nesta pesquisa foram calculados de acordo com a fórmula apresentada por Silva, Vasconcelos e Del-Moral-Hernandez (2021). Dessa forma, as dimensões obtidas foram de 5x5 para o primeiro mapa e 4x4 no segundo mapa, ambos com uma topologia hexagonal.

Após a geração do primeiro mapa, pode-se perceber através de um gráfico de distância média entre os objetos do mapa e os seus vizinhos mais próximos, quais são os neurônios no mapa que tem as menores distâncias (possuem sobreposição de dados). Dessa forma, mapearam-se os objetos com menos informatividade presentes nesses neurônios e formou-se um novo conjunto de dados contendo objetos sobrepostos.

O novo conjunto de dados foi separado em base de treinamento e de teste, de modo que 80% dos dados ficaram no conjunto de treinamento e 20% no conjunto de teste. O conjunto de treinamento foi usado para a geração de um segundo Mapa Auto-Organizável e após isso foi feita uma associação entre os objetos desse conjunto com os neurônios do mapa.

No conjunto de teste, um objeto foi escolhido e o neurônio mais próximo a ele foi observado. Em seguida, o algoritmo *SOMLI-KNN* pôde ser executado e foram realizados experimentos comparativos entre ele, o *LI-KNN* e o *KNN* com o intuito de verificar a acurácia dos classificadores em diferentes abordagens. Nos experimentos envolvendo o SOM, variou-se o número de neurônios mais próximos a serem considerados no conjunto de teste do modelo em um intervalo de um neurônio até quatro neurônios.

Para uma análise qualitativa, realizou-se um conjunto de experimentos com os resultados médios de acurácia dos modelos apresentados. Nesta análise, além do resultado médio da acurácia, também se apresentou o tempo de processamento de cada algoritmo.

Os experimentos foram mostrados de maneira subdividida. A princípio, foram apresentados os experimentos comparativos com destaque para uma análise comparativa e depois com experimentos envolvendo uma análise de tempo e acurácia. A abordagem que foi utilizada nos experimentos está definida no diagrama da Figura 1.

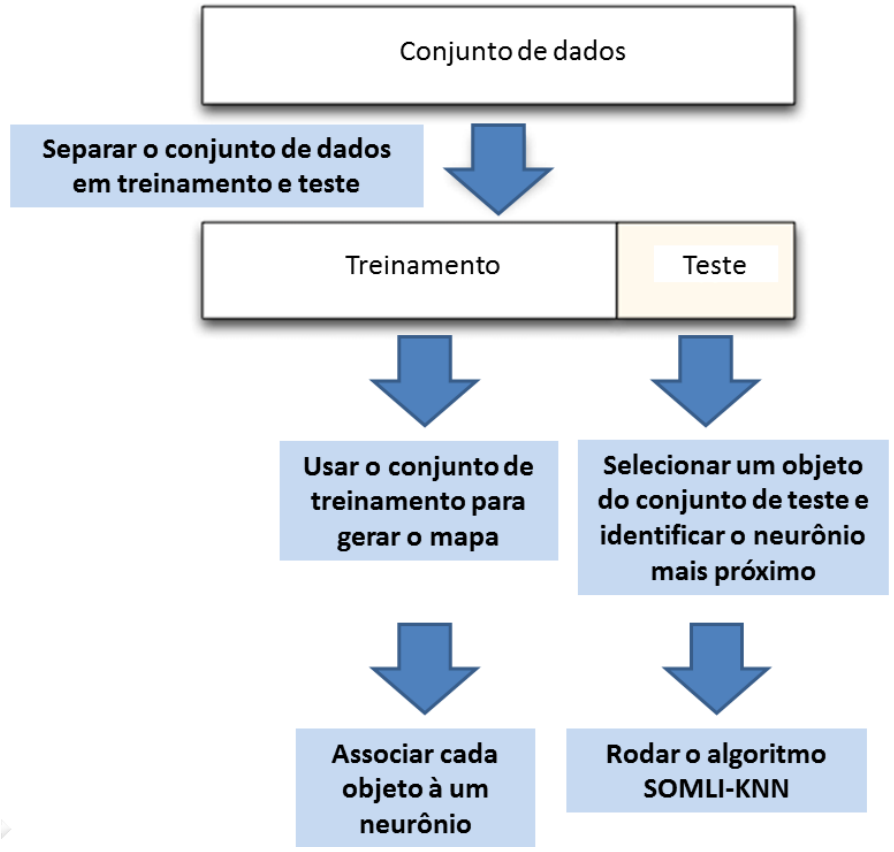

Figura 1- Diagrama *da abordagem metodológica utilizada na pesquisa.*

4. Resultado e discussão

4.1. Análise Comparativa

Na Figura 2 encontra-se a representação do primeiro mapa gerado. Ele possui dimensão 5*x*5 e topologia hexagonal. A utilização do SOM permite que se reduza o número de dimensões do conjunto de dados, facilitando a visualização dos dados. Pode-se observar neste mapa que os dados estão distribuídos nos 25 neurônios do mapa, de modo que alguns neurônios possuem apenas dados rotulados com uma das classes (0 ou 1), enquanto que outros neurônios possuem dados contendo apenas uma classe.

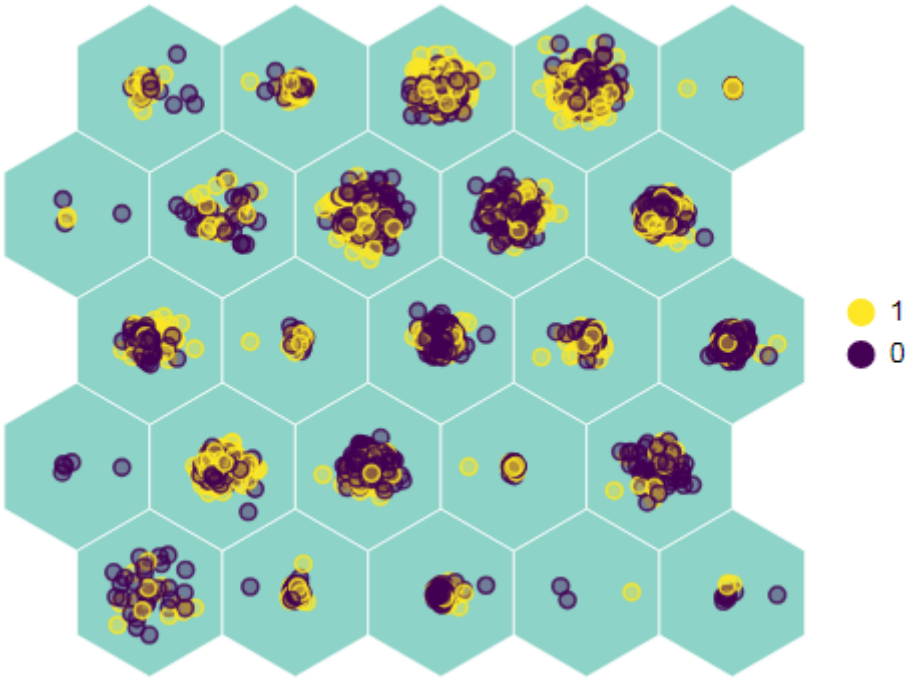

Figura 2 - Imagem do primeiro mapa gerado contendo os valores de 0 e 1 representando resultado positivo e negativo de Covid respectivamente.

A vantagem da abordagem utilizando SOM é que ela permite que seja feita a organização e manutenção topológica dos dados, de modo que é possível explorar apenas uma região de interesse ao invés do conjunto todo. Isso pode ser observado na Figura 3, onde os neurônios são classificados de acordo com a distância média entre seus vizinhos: uma cor mais clara indica que a distância é grande, enquanto que uma cor mais escura representa que a distância média é pequena.

Como foi observado por Grande e Silva (2019), o algoritmo *LI-KNN* possui um desempenho melhor em regiões sobrepostas. Sendo assim, os objetos presentes nos neurônios com menor distância (região roxa da Figura 3) foram mapeados e um novo conjunto de dados contendo objetos sobrepostos foi formado. Os experimentos descritos a seguir irão utilizar este novo conjunto de dados.

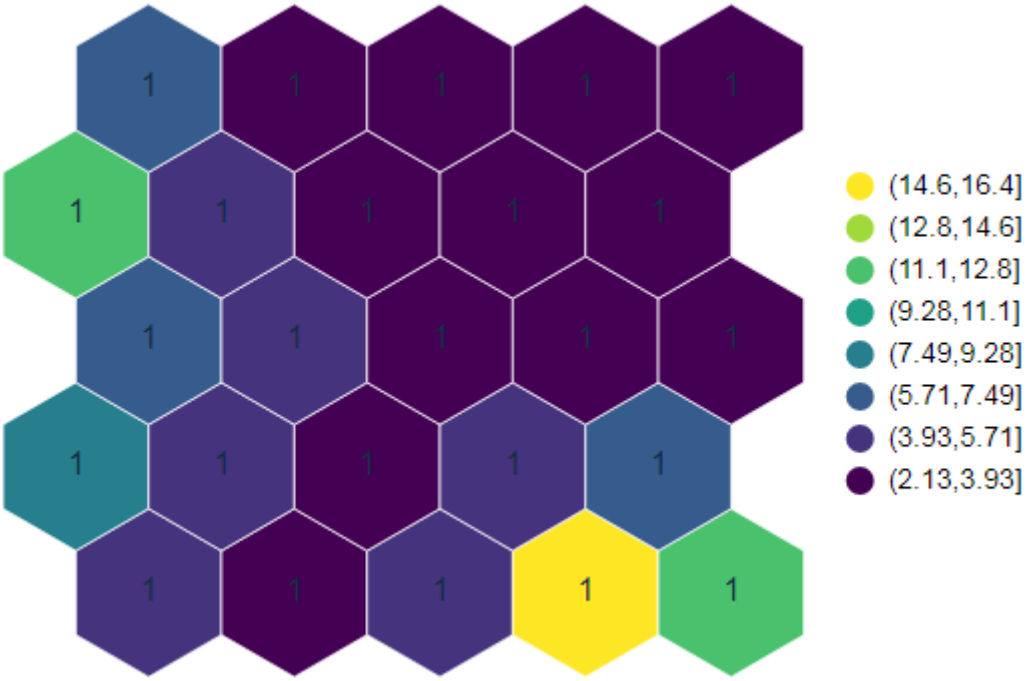

Figura 3 – Mapa auto-organizável com distância média entre os vizinhos

Após a criação do novo conjunto de dados, um segundo Mapa Auto-Organizável foi gerado. Este segundo mapa foi utilizado junto com o algoritmo *SOMLI-KNN* para fazer experimentos comparativos com o *LI-KNN* e o *KNN*.

Com o intuito de averiguar quantos neurônios mais próximos deveriam ser considerados no conjunto de teste da abordagem feita com o *SOMLI-KNN*, foi realizado um experimento comparativo em que o número de neurônios mais próximos a serem considerados no conjunto de teste do modelo foi variado em um intervalo de um neurônio até quatro neurônios. Os resultados desse experimento podem ser visualizados na Figura 4.

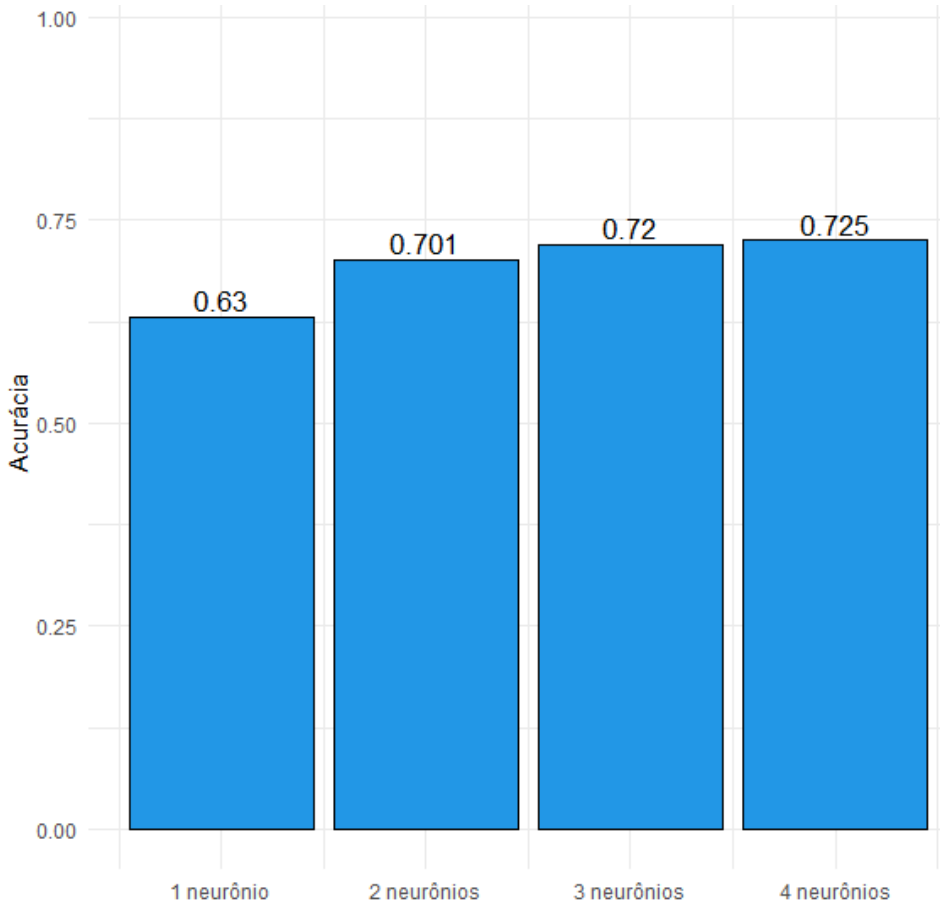

Figura 4 – Comparação das acurácias do *SOMLI-KNN* variando o número de neurônios próximos de um até quatro.

Analisando o experimento ilustrado na Figura 4, pode-se notar que o algoritmo *SOMLI-KNN* com quatro neurônios possui o melhor resultado. Sendo assim, para os estudos comparativos a seguir, a abordagem com o classificador *SOMLI-KNN* irá utilizar quatro neurônios mais próximos no conjunto de teste.

Com a parametrização dos neurônios definida no *SOMLI-KNN*, o algoritmo pôde ser executado. A abordagem principal utilizada neste algoritmo envolveu a separação do conjunto de dados em conjunto de treinamento e conjunto de teste. Após isso, o conjunto de treinamento foi usado para gerar o novo mapa e cada objeto desse conjunto foi associado a um neurônio presente no mapa. No conjunto de teste, um objeto foi selecionado e o neurônio mais próximo a ele foi identificado.

Com essas informações coletadas, o algoritmo *SOMLI-KNN* foi comparado com o *LI-KNN* e *KNN*. Os resultados podem ser observados na Tabela 1.

Tabela 1 – Comparação das métricas dos algoritmos KNN, LI-KNN e SOMLI-KNN

Algoritmos	Acurácia (Média)	Acurácia (Desvio padrão)	Tempo (Predição)
KNN	0.727	0.018	0.024
LI-KNN	*0.718*	0.016	5.400
SOMLI-KNN	0.723	0.018	0.530

4.2. Análise qualitativa

O objetivo desta seção é apresentar análises qualitativas dos resultados feitos de forma comparativa. Além do uso de Mapas Auto Organizáveis, o experimento também envolveu a comparação do *SOMLI-KNN* com outros algoritmos da literatura.

Os resultados apresentados na Tabela 1 representam a comparação em termos de acurácia (média e desvio padrão) e tempo de predição. Observa-se que o *SOMLI-KNN* possui um desempenho um pouco maior do que o *LI-KNN*, mostrando a importância do Mapa Auto Organizável nesta abordagem. O *KNN* e o *SOMLI-KNN* têm desempenhos semelhantes entre si, o que representa um bom resultado para o *SOMLI-KNN*, visto que o *KNN* é um algoritmo de classificação conhecido na literatura pela sua eficácia. Pode-se perceber que o tempo que leva-se para realizar a classificação com o *LI-KNN* é maior que todos os outros algoritmos.

5. Conclusão e Trabalhos Futuros

O artigo apresentou uma nova abordagem para realizar a classificação dos dados em exames médicos de pacientes com COVID-19 com a utilização de Mapas Auto Organizáveis e o algoritmo *SOMLI-KNN*. Os resultados foram comparados com outros classificadores da literatura como o *KNN* e o *LI-KNN*.

De acordo com os resultados obtidos, pode-se concluir que a utilização de *SOM* acelerou o treinamento do algoritmo *SOMLI-KNN*, permitindo que ele fosse executado de forma mais rápida que o *LI-KNN*. Além disso, o *SOMLI-KNN* obteve um desempenho mediano em um conjunto de dados com neurônios próximos ao objeto de teste.

Pretende-se futuramente usar uma abordagem do algoritmo *KNN* junto com *SOM* para verificar se o tempo de predição do *KNN* pode ser ainda mais reduzido a partir dessa técnica. Posteriormente, planeja-se também utilizar os protótipos do *SOM* para rotular os dados do *SOMLI-KNN* e comparar este resultado com outros algoritmos da literatura como *Hclust* e *K-Means*.

Referências

Avila E, Dorn M, Alho CS, Kahmann A (2020) Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios. arXiv preprint arXiv:2005.10227

Cabitza F, Campagner A, Ferrari D, Di Resta C, Ceriotti D, Sabetta E, Colombini A, De Vecchi E, Banfi G, Locatelli M et al (2020) Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. Clin Chem Lab Med (CCLM) 1(ahead-of-print)

Ciotti, M.; Angeletti, S.; Minieri M.; Giovannetti, M.; Benvenuto, D.; Pascarella, S.; Sagnelli, C.; Bianchi, M.; Bernardini, S.; Ciccozzi, M. COVID-19 outbreak: an overview. Chemotherapy (2019), 64, 215–223.

Cover, T; Hart, P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, Vol. 13, No.1, pp. 21-27, January (1967).

de Moraes Batista AF, Miraglia JL, Donato THR, Chiavegatto Filho ADP (2020) COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv

Fleury. Conheça os diferentes tipos de teste para COVID-19, 2020. Disponível em:<https://www.fleury.com.br/noticias/conheca-os-diferentes-tipos-de-teste-para-covid-19>. Acesso em: 12 Nov. (2021)

Grande, Vinícius Gomes Pajaro; Da Silva, Leandro Augusto. Análise de sensibilidade dos parâmetros do algoritmo k vizinhos informativos mais próximos para problemas de classificação de dados. In: XV Jornada de Iniciação Científica e IX Mostra de Iniciação Tecnológica-2019. (2019).

Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques, third edition Morgan Kaufmann Publishers

Kohonen, T. Essentials of the self-organizing map. Neural Networks (2013), 37, 52 – 65. 402 doi:https://doi.org/10.1016/j.neunet.2012.09.018.

Maimon, O; Rokach, L. Data Mining and Discovery Knowledge Handbook. 2nd edition, Springer (2010).

Ministério da Saúde. Sobre a doença, 2019. Conteúdos sobre o Coronavírus. Disponível em: <https://coronavirus.saude.gov.br/sobre-a-doenca>. Acesso em: 30 Nov. (2020)

Moreira L.J.; Silva, L.A. Prototype Generation Using Self-Organizing Maps for Informativeness-Based Classifier. Computational intelligence and neuroscience 2017 (2017).

Silva, Leandro A., de Vasconcelos, Bruno P., and Del-Moral-Hernandez, Emilio. 'A Model to Estimate the Self-Organizing Maps Grid Dimension for Prototype Generation'. 1 Jan. (2021) : 321 – 338.

Silva L. A., Del-Moral-Hernandez E. A SOM combined with KNN for classification task. Proceedings of the 2011 International Joint Conference on Neural Network, IJCNN 2011; August (2011); San Jose, Calif, USA. pp. 2368–2373.

Soares F, Villavicencio A, Fogliatto FS, Rigatto MHP, Anzanello MJ, Idiart M, Stevenson M (2020) A novel specific artificial intelligence-based method to identify COVID-19 cases using simple blood exams. medRxiv

Song Y., Huang J., Zhou D., Zha H., Giles C. L. IKNN: Informative K-Nearest Neighbor Pattern Classification. Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery; (2007); Berlin, Germany. Springer; pp. 248–264.

Souza, A.A.d., Almeida, D.C.d., Barcelos, T.S. et al. Simple hemogram to support the decision-making of COVID-19 diagnosis using clusters analysis with self-organizing maps neural network. Soft Comput (2021). https://doi.org/10.1007/s00500-021-05810-5

Organização Pan-Americana da Saúde. Histórico da pandemia de COVID-19, 2020. Disponível em:<https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19>. Acesso em: 12 Nov. (2021)

Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. JAMA. (2020) Feb;323((11)):1061

Wettschereck, D. A Study of Distance-Based Machine Learning Algorithms. Doctor of Philosophy Dissertation. Oregon State University (1994).

Wu, X. and Kumar, V. , The top ten algorithms in data mining (CRC Press) (2009)