Seleção de Características Biológicas para Prognóstico de Câncer: Revisão Sistemática da Literatura

Selection of Biological Characteristics for Cancer Prognosis: Systematic Literature Review

Selección de características para pronóstico del cáncer: revisión sistemática de la literatura

Maxwell E. A. Silva[1], Victor G.L. Holanda[1], Rodrigo S. da Silva[1], Paulo V.L. Severiano[1], Rafael de Amorim Silva[1]

1 Universidade Federal de Alagoas - Instituto de Computação - Centro de Pesquisa em Tecnologias Emergentes - Maceió (AL), Brasil.

Autor correspondente: Maxwell E. A. Silva
E-mail: meas@ic.ufal.br

Resumo

Este trabalho elabora uma revisão sistemática sobre seleção de características em prognóstico médico. O objetivo principal é compreender os principais métodos de seleção utilizados em modelos de aprendizagem de máquina para predizer doenças como câncer. As seguintes fontes eletrônicas de dados foram utilizadas: IEEE, ACM, Elsevier, Springer e PubMed. Utilizou-se artigos científicos primários de periódicos e conferências escritos na língua inglesa como critérios principais de elegibilidade. O protocolo desta revisão selecionou 21 artigos utilizando o método PICOS como critério de qualidade. Os resultados dessa revisão destacam o método ReliefF como um dos mais eficientes métodos de seleção de características, reforçando o seu uso em situações como o uso de microarranjos de expressão gênica das células de câncer.
Descritores: Seleção de características; Métodos; Prognóstico Câncer;

Abstract

This work elaborates a systematic review on the selection of characteristics in medical prognosis. The main objective is to understand the main selection methods used in machine learning models to predict diseases such as cancer. The following electronic data sources were used: IEEE, ACM, Elsevier, Springer and PubMed. Primary scientific

articles from journals and conferences written in the English language were used as the main eligibility criteria. The review protocol selected 21 articles using the PICOS method as a quality criterion. The results of this review highlight the ReliefF method as one of the most efficient methods for selecting characteristics, reinforcing its use in situations such as the use of microarrays of gene expression of cancer cells.
Keywords: Features selection; Methods; Cancer Prognosis;

Resumen

Este trabajo elabora una revisión sistemática sobre la selección de características en el pronóstico médico. El objetivo principal es comprender los principales métodos de selección utilizados en los modelos de aprendizaje automático para predecir enfermedades como el cáncer. Se utilizaron las siguientes fuentes de datos electrónicos: IEEE, ACM, Elsevier, Springer y PubMed. Los principales criterios de elegibilidad fueron artículos científicos primarios de revistas y congresos escritos en inglés. El protocolo de revisión seleccionó 21 artículos utilizando el método PICOS como criterio de calidad. Los resultados de esta revisión destacan el método ReliefF como uno de los métodos más eficientes para seleccionar características, reforzando su uso en situaciones como el uso de microarrays de expresión génica de células cancerosas.
Keywords: Selección de características; Métodos; Pronóstico Cáncer;

Introdução

No contexto de aprendizagem de máquina (*Machine Learning* ou ML), características podem ser consideradas propriedades de um fenômeno observado. Estas são usualmente utilizadas em um processo de reconhecimento sistemático de padrões, através da aplicação de modelos de ML. A seleção de características, ou redução de dimensionalidade, pode ser denominada como o processo que visa a redução da quantidade de características que são utilizadas em um modelo. Esta redução pode ser realizada através de processamento e análise destas características, com o objetivo de encontrar quais destas são mais significativas dentre todas as propriedades existentes [1]. Os métodos de seleção de características geralmente são

divididos em três tipos: (i) *filters*; (ii) *wrappers*; e (iii) *embedded*. *Filters* são utilizados na fase de pré-processamento dos dados, não têm relação com o modelo de aprendizagem utilizado. Já os *wrappers* são usados na fase de treinamento do modelo preditivo, por esta razão podem ser menos performáticos que os *filters* do ponto de vista computacional. Por fim, temos os métodos *embeddeds* que podem ser caracterizados como uma combinação dos dois métodos citados anteriormente [2].

Na aplicação de modelos de ML no contexto de prognóstico de câncer, recentemente a seleção de características se tornou uma ferramenta fundamental nas etapas de pré-processamento dos dados, com um alto número de características como conjuntos de dados que contêm microarranjos de expressão gênica. Pois este tipo de conjunto de dados é composto de centenas de milhares de características, com um número pequeno da amostra analisada. Algoritmos de ML geralmente não trabalham muito bem com este tipo de dado. Portanto, faz-se necessário a redução do número de características para remover de propriedades redundantes e irrelevantes, e para que o algoritmo utilizado possa apresentar um bom resultado [3].

Portanto, este trabalho apresenta uma revisão sistemática que compreenda os principais métodos de seleção de características utilizados em modelos de ML para predizer doenças como câncer. Utilizou-se um protocolo baseado na recomendação PRISMA e no trabalho de Kitchenham e Charters. Fontes eletrônicas de dados como IEEE, ACM, Elsevier, Springer e PubMed foram incluídas e questões de pesquisa foram definidas para nortear a estratégia de busca dos artigos na literatura. O protocolo desta revisão selecionou 21 artigos utilizando critérios de qualidade baseado na estratégia PICOS [4]. Entre os achados científicos, identificou-se o método ReliefF como um dos mais eficientes para selecionar características. A estrutura do trabalho é definida da seguinte maneira: A Seção ‘Método’ apresenta o protocolo desenvolvido por esta revisão. A Seção ‘Resultados’ ilustra os resultados obtidos pela execução do protocolo. A Seção ‘Discussão’ interpreta os principais achados científicos encontrados nesta revisão. A Seção ‘Conclusão’ apresenta as considerações finais deste artigo.

Método

Esta revisão sistemática considera os 27 itens da recomendação PRISMA [5] e se

baseia em partes no protocolo definido por Kitchenham e Charters [6]. Para a sistematização dessa revisão, utilizaram-se as seguintes ferramentas: (i) Mendeley, uma ferramenta oferecida pela ELSEVIER para organização de artigos científicos, sendo utilizada na etapa de seleção dos estudos investigados nesta revisão; e (ii) Google Sheets, ferramenta utilizada para organizar e sintetizar os achados da pesquisa.

Protocolo Utilizado

O protocolo utilizado nesta revisão.consiste nos seguintes elementos: (i) elaboração das questões de pesquisa; (ii) definição das palavras chaves; (iii) escolha das fontes científicas; (iv) definição dos critérios de inclusão e exclusão dos artigos relacionados nesta revisão; e (v) estratégia utilizada na busca dos artigos investigados neste trabalho. Tais elementos foram escolhidos por esta revisão no intuito de sistematicamente selecionarmos os artigos mais relevantes sobre o uso de métodos relacionados a seleção de características biológicas para aumentar a acurácia no prognóstico do câncer.

Questões de Pesquisa e Palavras-Chave

Uma revisão sistemática deve-se basear em questionamentos que norteiem a busca por informação nos artigos investigados para responder apropriadamente cada questão levantada. Neste sentido, a Tabela 1 identifica as principais questões de pesquisa utilizadas nesta revisão sistemática.

A questão de pesquisa principal (QP) enfatiza a necessidade de se conhecer a relevância dos métodos de seleção de características para realizar um prognóstico médico eficiente no tratamento de pacientes com câncer. A questão QS1 tenta identificar em quais tipos de câncer são aplicados o processo de seleção de características para reduzir as características irrelevantes nos dados do paciente e escolher aquelas com maior impacto na predição acurada da condição clínica do paciente. A questão QS2 tenta identificar quais tipos de dados do paciente são utilizados no processo de seleção de características para se obter um adequado desempenho nos modelos de aprendizado de máquina. A questão QS3 tenta identificar

quais métodos devem ser utilizados no processo de seleção de características para se

Tabela 1 – Questões levantadas sobre seleção de características para prognóstico de câncer

Questão Principal
● ***QP: Como a Seleção de Características tem sido aplicada nos modelos de aprendizado de máquina para prognóstico de câncer?***
Questões Secundárias
● ***QS1: Quais os principais tipos de câncer em que são utilizados métodos de seleção de características nos modelos de aprendizado de máquina?***
● ***QS2: Quais são os tipos de dados em que são utilizados métodos de seleção de características nos modelos de aprendizado de máquina?***
● ***QS3: Quais são os principais métodos de seleção de características aplicados nos modelos de aprendizado de máquina?***
● ***QS4: Qual é a acurácia do modelo de aprendizado após aplicação dos métodos de seleção de características no prognóstico de câncer?***

obter uma boa performance nos modelos de aprendizado de máquina. A questão QS4 tenta identificar qual é a performance do modelo preditivo após a aplicação de um método de seleção de características. A partir das questões de pesquisa levantadas, entende-se que as palavras principais que norteiam esta revisão são: (i) Seleção de características; (ii) Métodos; (iii) Prognóstico Câncer;

Critérios de elegibilidade

Para definir um patamar de qualidade na seleção dos artigos investigados neste trabalho, definem-se três tipos de critérios de elegibilidade: (i) critérios de inclusão; (ii) critérios de exclusão; e (iii) critérios de qualidade.

Nos critérios de inclusão, são considerados apenas trabalhos científicos primários publicados em periódicos ou em anais de eventos escritos na língua inglesa. Os artigos devem ser relacionados exclusivamente ao uso de seleção de características para prognóstico em câncer, desconsiderando os artigos que utilizam predição para outras doenças ou para fins de diagnóstico clínico. São considerados apenas artigos publicados nos anos de 2015 a 2020 e que já tenham sido publicados eletronicamente ou de forma impressa.

Nos critérios de exclusão, eliminam-se os estudos que não atendem aos critérios de inclusão, como artigos secundários (*surveys*, revisões sistemáticas) e terciários (revisões de revisões sistemáticas), artigos redigidos em outros idiomas, artigos

redundantes, duplicados ou indisponíveis. Também são excluídos aqueles artigos que investigam apenas aspectos do prognóstico ou seleção de características em outros contextos. Por último, cada estudo deve atender um limiar mínimo de qualidade para ser aceito nesta revisão.

Nos critérios de qualidade, define-se tal limiar alcançado por um esquema de pontuação que considera a estratégia PICOS nos estudos investigados. Esta estratégia consiste em analisar os trabalhos selecionados para identificar se todos incluem referências sobre a população, os pacientes ou o problema, além do método de intervenção e comparação, o resultado obtido e o projeto do estudo apresentado. Cada Critério PICOS é analisado pelos autores deste trabalho e uma nota de 0 a 1 (i.e. 1 se o critério for atendido, 0.5 se for parcialmente atendido e 0 se o critério não for atendido) é dada para cada critério de qualidade, elaborando-se um score de pontos obtidos pela somatória desses critérios. Os trabalhos que não obtiverem uma pontuação mínima de 3 pontos será excluído do processo de seleção.

Fontes de informação

A investigação de aspectos relacionados à inteligência artificial aplicada no contexto de saúde requer um conhecimento multidisciplinar. Portanto, os autores têm considerado esta multidisciplinaridade e têm escolhido fontes de pesquisa tanto na área pura de computação quanto na área de informática aplicada à saúde. Desta forma, foram escolhidas as seguintes fontes de informação: (i) ACM; (ii) PubMed; (iii) IEEE; (iv) Springer; e (v) Elsevier (através do *Science Direct*). A execução deste protocolo foi realizada no dia 25 de setembro de 2020.

Busca

Para realizar esta revisão sistemática, utilizou-se apenas bases eletrônicas em motores de busca da *World Wide Web*. A *string* de busca foi elaborada considerando as palavras chaves retiradas das questões de pesquisa. A seguinte *string* de busca foi utilizada:

("machine learning") AND (("feature selection") OR ("features selection")) AND

((method) OR (algorithm) OR (technique)) AND (("cancer prognosis") OR ("cancer prediction"))

Os filtros dos motores de busca das fontes de informação foram configurados para considerar os critérios de inclusão e exclusão. Na maioria dos motores de busca, escolheu-se a opção por busca avançada para filtrar os critérios. A mesma *string* de busca foi utilizada em todas as fontes de pesquisa, sendo colocado parênteses para associar cada palavra-chave com seu adjacente e apóstrofos para incluir a expressão como uma palavra única. Assim, conseguiu-se coletar uma larga quantidade de estudos diretamente relacionados às nossas questões de pesquisa.

2.5 Seleção dos estudos

O protocolo foi executado de acordo com as configurações apresentadas acima e a seleção de estudos ocorreu em 6 etapas. Na etapa 1, a *string* de busca é aplicada nos motores de busca das fontes de informação, coletando 301 artigos destas fontes. Na etapa 2, utilizou-se os filtros existentes nestas fontes de pesquisa para coletar apenas os estudos publicados em anais e periódicos e definir o intervalo de tempo de publicação de acordo com o critério de inclusão. Assim, estudos não revisado por pares como enciclopédias, normas, cursos, resenhas, livros, entre outros, foram excluídos no processo. A etapa 3 corresponde a exclusão de todos os artigos duplicados, redundantes e indisponíveis. Na etapa 4, realizou-se uma revisão dos títulos, das palavras-chave, do local de publicação e do resumo para excluir os artigos que não atendessem aos critérios de inclusão e exclusão. Na etapa 5, todos os artigos incluídos na etapa 4 foram recuperados de suas respectivas fontes de informação e lidos. Na etapa 6, foi feita uma avaliação da qualidade dos artigos considerando a inclusão de elementos relacionados a abordagem PICOS. Uma pontuação foi definida por artigo e foram selecionados apenas aqueles que atenderam a pontuação mínima de 3 pontos. A Tabela 2 relaciona os estudos selecionados, suas referências, título e descrição.

Coleta e Lista de dados

Os dados utilizados nesta revisão sistemática foram extraídos dos artigos selecionados e inseridos em uma planilha eletrônica para fins de análise e geração de gráficos. Os

dados de cada artigo selecionado foi conferido duas vezes para garantir a qualidade dos trabalhos escolhidos. Coletaram-se informações sobre os tipos de métodos investigados, em quais tipos de câncer é utilizada a seleção de características e em quais tipos de dados esta abordagem é usada. Todos os resultados obtidos nesta revisão são apresentados por meio das seguintes ferramentas: (i) tabelas; e (ii) gráfico de linhas.

Resultados

Seleção de estudos

O processo de obtenção sistemática de trabalhos que correspondem as questões de pesquisa obteve os seguintes resultados: na etapa 1, um total de **628** artigos foram coletados pelas fontes de informação. Em seguida (na etapa 2), **301** artigos foram

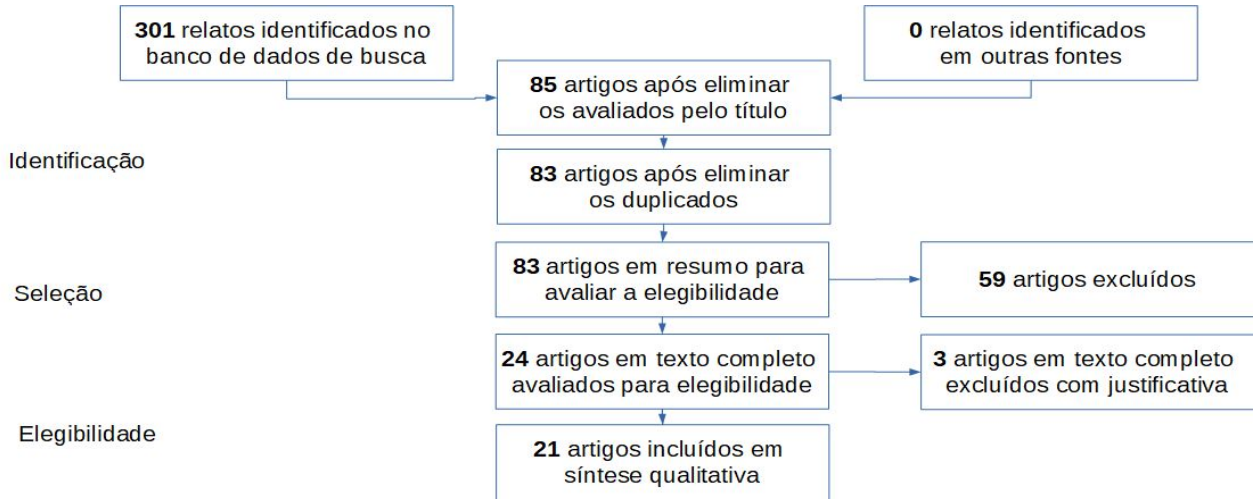

Figura 1 – Fluxograma que representa a execução do protocolo desta revisão

incluídos após a aplicação dos filtros, sendo **327** artigos eliminados por não serem artigos primários, ou por não pertencerem a periódicos ou conferências. Na etapa 3, foram removidos os artigos cujo título não estava de acordo com o trabalho proposto, neste caso foram excluídos **216** artigos. Já na etapa 4, foram removidos **2** artigos duplicados. Na etapa 5, **24** artigos foram incluídos após análise dos seus respectivos resumos. Por fim na etapa 6, foram removidos **3** artigos completos cujo objetivo não

estava de acordo com a proposta deste trabalho. Ao final das etapas anteriores, restaram **21** artigos completos para serem analisados qualitativamente de acordo com os critérios da análise qualitativa de todos os artigos incluídos considerando a estratégia PICOS, porém nenhum artigo foi excluído por não atingir a nota mínima definida para estes critérios, ou seja, a nota mínima de 3 pontos.

Características dos estudos

Nesta revisão, os seguintes atributos foram considerados nos artigos selecionados na etapa 6: (i) informações de metadados dos veículos científicos, tais como o nome da editora e o nome da conferência ou periódico em questão; (ii) ano de publicação do artigo em questão; (iii) tipos de câncer (para qual tipo de câncer a pesquisa direciona seus resultados, por exemplo, câncer de pulmão, mama, etc); (iv) tipos de dados utilizados para fazer o prognóstico do câncer (v) contribuições do artigo (e.g. um método, uma técnica, uma abordagem, uma ferramenta, etc); e (vi) *score* do método PICOS.

Quadro 1 – Metadados dos Artigos Selecionados [6]

ID	Autores e ano	Título do artigo	Ano de Pub.	Local de Pub.	Editora
T1	Ghaisani, Fakhirah D; Wasito, Ito; Faturrahman,Moh; Mufidah, Ratna.	Deep Belief Networks and Bayesian Networks for Prognosis of Acute Lymphoblastic Leukemia	2017	International Conference Proceeding Series (ICPS)	ACM
T2	Abdelaziz, Esraa H; Kamal, Sanaa M; El-Bhanasy, Khaled; Ismail, Rasha.	The Application of Data Mining Techniques and Feature Selection Methods in the Risk Classification of Egyptian Liver Cancer Patients Using Clinical and Genetic Data	2019	International Conference on Software and Information Engineering	ACM
T3	Xia, Chao; Xiao, Yawen; Wu, Jun; Zhao, Xiaodong; Li, Hua.	A Convolutional Neural Network Based Ensemble Method for Cancer Prediction Using DNA Methylation Data	2019	International Conference on Machine Learning and Computing	ACM
T4	Singh, Bikesh Kumar.	Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm	2019	Biocybernetics and Biomedical Engineering	Elsevier
T5	Waseem, M H; Nadeem,M S A; Abbas, A; Shaheen, A;Aziz, W;	On the Feature Selection Methods and Reject Option Classifiers for Robust Cancer Prediction	2019	IEEE Access	IEEE

	Anjum, A; Manzoor, U; Balubaid, M A; Shim, S.				
T6	Nithya, B; Iango, V	Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction	2019	SN Applied Sciences	Springer
T7	Dhanya, R; Paul, Irene Rose; Akula, Sai Sindhu Sivakumar, Madhumathi Nair, Jyothisha J.	F-test feature selection in Stacking ensemble model for breast cancer prediction	2020	Procedia Computer Science	Elsevier
T8	Shafi, A S M Molla, M M Imran Jui, Julakha Jahan Rahman, Mohammad Motiur.	Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques	2020	SN Applied Sciences	Springer
T9	Khourdifi, Y Bahaj, M.	Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms	2018	Int. Symp.on Adv. Electrical and Comm. Tech. (ISAECT)	IEEE
T10	Thara, L Gunasundari, R.	Adaptive feature selection method based on particle swarm optimization for gastric cancer prediction	2017	Int. Conf. on Comm. and Electro. Systems (ICCES)	IEEE
T11	Guixia Kang Zhuang Ni.	Research on early risk predictive model and discriminative feature selection of cancer based on real-world routine physical examination data	2016	IEEE International Conference on Bioinformatics and Biomedicine (BIBM)	IEEE
T12	Agarwalla, Prativa Mukhopadhyay, Sumitra	Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach	2018	Applied Soft Computing	Elsevier
T13	Mourad, Moustafa Moubayed, Sami Dez., Aaron; Mourad, Youssef; Park, Kyle; Torreblanca-Zanca; Alb.; Torrecilla, José S; Cancilla, J.C Wang,Jiwu	Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis.	2020	Scientific reports	PubMed
T14	Santhakumar, D Logeswari, S	Efficient attribute selection technique for leukaemia prediction using microarray gene data	2020	Soft Computing	Springer
T15	Azzawi, H Hou, J Xiang, Y Alanni, R	Lung cancer prediction from microarray data by gene expression programming	2016	IET Systems Biology	IEEE
T16	Mei, X	Predicting five-year overall survival in patients with non-small cell lung cancer by reliefF algorithm and random forests	2017	IEEE 2nd Adv. Info Tech., Electronic and Automation Ctrll Conf. (IAEAC)	IEEE
T17	Zhang, D; Zou, L; Zhou, X; He, F	Integrating Feature Selection and Feature Extraction Methods With	2018	IEEE Access	IEEE

		Deep Learning to Predict Clinical Outcome of Breast Cancer			
T18	Shen, Y; Wu, C; Liu, C; Wu, Y; Xiong, N	Oriented Feature Selection SVM Applied to Cancer Prediction in Precision Medicine	2018	IEEE Access	IEEE
T19	Shanthi, S; Rajkumar, N	Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods	2020	Neural Processing Letters	Springer
T20	Ke, W; Wu, C;Wu, Y Xiong, N N	A New Filter Feature Selection Based on Criteria Fusion for Gene Microarray Data	2018	IEEE Access	IEEE
T21	Doreswamy Salma, M U	PSO based fast K-means algorithm for feature selection from high dimensional medical data set	2016	Int. Conf. on Intelligent Systems and Control (ISCO)	IEEE

O item (i) representa a localização e tema das conferências e periódicos considerados. Nesta revisão a associação científica IEEE obteve o maior número de artigos recuperados após a aplicação da string de busca (10 artigos, sendo 47,6% do total das fontes de pesquisa), seguido da Springer (4 artigos, representando 19% do total), ACM (4 artigos, representando 14,3% do total), ELSEVIER (4 artigos, representando 14,3% do total) e por último a associação científica PubMed (1 artigo, representando 4,8% do total). O item (ii) consiste nos dados sobre o ano de publicação dos estudos selecionados. (i) 2016 (3 estudos, 14,3%); (ii) 2017 (3 estudos, 14,3%); (iii) 2018 (5 estudos, 23,8%); (iv) 2019 (5 estudos, 14,3%); (v) 2020 (5 estudos, 23,8%, o que demonstra uma crescente preocupação nos últimos anos sobre a seleção de características no contexto de prognóstico de câncer. O item (iii) consiste nos dados sobre os tipos de câncer que são analisados nos estudos selecionados. (i) câncer na mama (5 estudos, 20%); (ii) câncer no pulmão (5 estudos, 20%); (iii) câncer em geral (4 studies, 16%); (iv) leucemia (3 estudos, 12%); (v) câncer no fígado (2 estudos, 8%); (vi) câncer no colo (2 estudos, 8%); (vii) câncer no rim (1 estudo, 4%); (viii) câncer no intestino (1 estudo, 4%); (ix) câncer cervical (1 estudo, 4%), estes dados demonstram uma maior preocupação nos últimos anos sobre a seleção de características no contexto de prognóstico de câncer sobre o câncer de mama e pulmão. Já o item (iv) consiste nos tipos de dados utilizados para o prognóstico. (i) dados do tumor (2 estudos, 8.7%); (ii) dados clínicos dos pacientes (8 estudos, 34.8%); e (iii) dados moleculares do câncer (13 estudos, 56.5%, portanto percebe-se que existe um maior

interesse em analisar fatores moleculares do câncer para melhorar o prognóstico desta patologia, podemos observar estes dados através da Figura 2). O item (v) consiste nos dados sobre as contribuições do artigo. (i) abordagem (14 estudos, 66%); (ii) técnica (6 estudos, 28%); (iii) modelo (1 estudo, 4%); logo os dados demonstram uma crescente preocupação nos últimos anos sobre análise dos métodos de seleção de características já existentes aplicados no contexto de prognóstico de câncer. O item (vi) consiste nos dados sobre o score obtido por cada artigo sobre a avaliação dos aspectos abordados na metodologia PICOS. (i) score 5 (7 estudos, 33.3%); (ii) score 4.5 (8 estudos, 38.1%); (iii) score 4 (6 estudos, 28.6%, o que demonstra que os artigos apresentam uma boa avaliação sobre os critérios definidos na metodologia PICOS sobre a seleção de características no contexto de prognóstico de câncer.

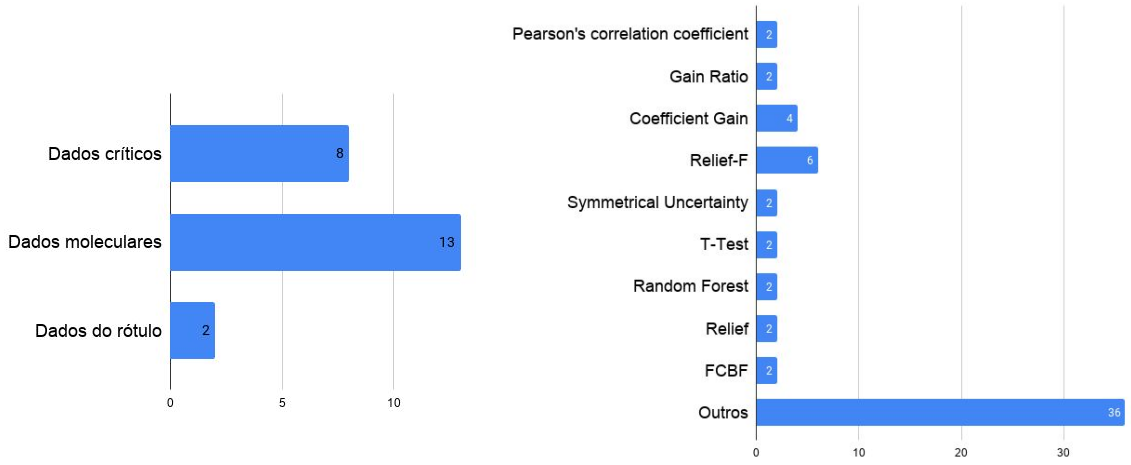

Figura 2 – (a) Tipos de dados utilizados para fazer o prognóstico de câncer; (b) Gráfico de linhas dos algoritmos mais utilizados

Síntese dos resultados

Estratégia PICOS os trabalhos (T1, T6, T16, T17, T18, T19 e T20) obtiveram nota 5 no score PICOS porque satisfizeram todos os aspectos analisados na estratégia PICOS. Já os trabalhos(T3, T7, T8, T10, T11, T12, T14 e T21) obtiveram nota 4.5 no score PICOS pois satisfizeram parcialmente os aspectos “população” ou “comparação” da estratégia utilizada como parâmetro. Por fim, os trabalhos (T2, T4, T5, T9, T13 e T15) obtiveram a nota 4 no score PICOS, estes obtiveram esta nota pois cumpriram parcialmente os requisitos “população” e “comparação” da estratégia PICOS.

Discussão

Sumário da evidência

Em resumo, a maioria dos trabalhos selecionados aplicaram a técnica ReliefF para seleção de características, neste estudo as técnicas que foram agrupadas as técnicas utilizadas apenas uma vez em um conjunto chamado de outros. Podemos observar as técnicas mais usadas na Figura 3. Esta escolha é devido este algoritmo ser considerado um filtro, ou seja, a seleção de características é feita independente do modelo de aprendizagem aplicado, resultando em uma boa performance computacional. Além disso, esta classe de algoritmo é capaz de fazer uma boa avaliação sobre a dependência das características dos dados. Esta boa avaliação dar-se pelo fato dele não utilizar o conceito de combinação de características para identificar uma possível dependência entre elas, mas sim usa a abordagem de vizinhos mais próximos para derivar estatísticas das características que possam explicar indiretamente a relação entre estas [7]. Por conta desta boa capacidade de avaliar a relação entre características, grupos de dados que contêm uma alta dimensionalidade se beneficiam, que é o caso de dados médicos que contêm informações de expressões gênicas, bastante utilizadas ultimamente para fazer o prognóstico da doença de câncer.

Limitações

Como ameaça a validade da nossa revisão, os seguintes pontos são considerados: (i) Esta pesquisa não considera doentes além de câncer; (ii) Os trabalhos que não estão escritos na língua inglesa não são considerados; (iii) Esta pesquisa não considera trabalhos que não estejam relacionados as aspectos de prognóstico da doença; (iv) Relatos clínicos podem ser encontrados em fontes não listadas, como literatura Grey, etc; (v) Viés do trabalho é encontrado em um nível moderado.

Conclusão

Este trabalho conduziu uma revisão sistemática para identificar quais são os métodos de seleção de características aplicados no processamento de dados relacionados a utilização de modelos de aprendizagem para prognóstico da doença câncer. Um

protocolo baseado na recomendação PRISMA foi definido e a estratégia PICOS foi utilizada para avaliar a qualidade dos trabalhos selecionados pelo protocolo. Os seguintes resultados foram obtidos e investigados: (i) entender sobre a importância de fazer a seleção de características quando trabalhamos com dados que tem um número muito alto de propriedades; (ii) entender quais são os tipos de câncer que vêm sendo aplicados a seleção de características como forma de melhorar a predição da condição clínica futura do paciente; (iii) relatar o método ReliefF como o mais adequado para base de dados com alta dimensionalidade.

Referências

1. Sherer, Tim. Feature Selection (Data Mining) [Internet]. Microsoft Documentation; 2018 [cited 2020 Ago 30]. Available from: https://docs.microsoft.com/en-us/analysis-services/data-mining/feature-selection-data-mining?view=asallproducts-allversions#:~:text=In%20this%20article&text=Feature%20selection%20refers%20to%20the,or%20features%20from%20existing%20data.
2. J. C. Ang, A. Mirzal, H. Haron and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 5, pp. 971-989, 1 September 2016, doi: 10.1109/TCBB.2015.2478454.
3. J. C. Ang, A. Mirzal, H. Haron and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 5, pp. 971-989, 1 September 2016, doi: 10.1109/TCBB.2015.2478454.
4. Methley, Abigail M et al. “PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews.” BMC health serv. research vol. 14 579. 21 Nov.2014,doi:10.1186/s12913-014-0579-0
5. Principais itens para relatar Revisões sistemáticas e Meta-análises: A recomendação PRISMA. Epidemiol. Serv. Saúde [Internet]. 2015 June [cited 2020 Sep 30] ; 24(2): 335-342. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2237-96222015000200335&lng=en. http://dx.doi.org/10.5123/S1679-49742015000200017
6. KITCHENHAM, Barbara; CHARTERS, Stuart. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 2007
7. Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics, 85, 189–203. https://doi.org/https://doi.org/10.1016/j.jbi.2018.07.014