Clusterização de Dados Abertos em Oncologia Usando Técnicas de Aprendizado de Máquina: um estudo preliminar sobre recidiva de câncer de próstata

Pedro Ferreira Crocco[1,2], André Eidi Maeda[1,2], Guilherme Cesar Soares Ruppert[1], Mariangela Dametto[1], Rodrigo Bonacin[1]

pcrocco@cti.gov.br, amaeda@cti.gov.br,
guilherme.ruppert@cti.gov.br,mdametto@cti.gov.br,
rodrigo.bonacin@cti.gov.br

[1] DIMEC, Centro de Tecnologia da Informação Renato Archer – CTI

[2] Universidade Estadual de Campinas - Unicamp

Abstract. *Due to the large amount of data currently produced by health institutions, data science can significantly contribute to the advancement of research, diagnosis and medical treatments. In this context, this article presents a study on the application of machine learning techniques for clustering using open databases in the field of oncology. Preliminary results with data from the database of Fundação Oncocentro de São Paulo for prostate cancer are presented and discussed in this work.*

Resumo. *Em função da quantidade de dados atualmente produzida pelas instituições de saúde, a ciência de dados pode contribuir de maneira significativa para o avanço em pesquisa, diagnóstico e tratamentos médicos. Neste contexto, este artigo apresenta estudo sobre a aplicação de técnicas de aprendizado de máquina para clusterização em base de dados abertas no domínio da oncologia. Resultados preliminares com dados da base da Fundação Oncocentro de São Paulo para câncer de próstata são apresentados e discutidos neste trabalho.*

Palavras-chaves: Aprendizado de Máquina, Clusterização, Inteligência Artificial, Informática em Saúde

1. Introdução

A Ciência de Dados em saúde pode ser entendida como um conjunto de técnicas utilizadas para extrair informações e auxiliar em conclusões a partir de conjuntos de dados, sejam estes dados demográficos, pessoais ou clínicos. A indústria da saúde produz uma enorme quantidade de dados, e o crescimento e aperfeiçoamento da tecnologia e das técnicas de aprendizado de máquina podem contribuir de forma impactante no desenvolvimento da medicina. Entre as contribuições possíveis estão as descobertas de novos remédios, predição de doenças por histórico do paciente, diagnóstico mais efetivo a partir de sintomas e tratamento personalizado [Panesar 2019].

Este artigo apresenta estudo realizado utilizando-se base de dados aberta no domínio da oncologia. Espera-se encontrar padrões nos dados de modo a auxiliar profissionais de saúde a encontrar e analisar fatores relacionados à recidiva de câncer, entre outros fatores de interesse médico. Para tanto, foram estudadas e aplicadas técnicas de aprendizado de máquina que permitem agrupar pacientes com características semelhantes, conhecidas como técnicas de clusterização.

A biblioteca scikit-learn[1] utilizada como base para foi a implementação. A scikit-learn é uma biblioteca desenvolvida na linguagem Python e possui código aberto. Ela inclui um conjunto amplo de classes e métodos para apoiar os desenvolvedores de aplicações práticas utilizando aprendizado de máquina.

O objetivo central dessa pesquisa é analisar a estrutura da base de dados da Fundação Oncocentro para os casos de câncer de próstata, aplicar o algoritmo selecionado após manipular as *features* de maior relevância e observar os *clusters* formados para tentar identificar algum padrão nos casos, especialmente naqueles em que a recidiva ocorreu.

O restante deste artigo está estruturado da seguinte forma: a seção 2 apresenta os conceitos utilizados ao longo da pesquisa, apresentando a base de dados e o algoritmo escolhido, a seção 3 apresenta a metodologia aplicada a partir dos conceitos já introduzidos, a seção 4 apresenta os resultados e a interpretação e, por fim, a seção 5 apresenta a conclusão tomada a partir dos dados coletados.

2. Conceitos, Técnicas, Algoritmos e Base de Dados

As técnicas de aprendizado de máquina podem ser divididas entre técnicas de predição, de classificação e de clusterização. A clusterização pode ser definida como a abordagem de dividir um conjunto de dados em grupos (*clusters*). Um critério comumente utilizado para esse fim é o da distância entre pontos que representam os valores das características (*features*) [Landau et al. 2011]. Dentro de um cluster, o ideal é que os pontos estejam o mais próximo possível uns dos outros e o mais distante possível dos pontos de outros clusters. Para tanto, cada algoritmo usa um critério de distância específico para lidar com diferentes tipos de dado.

Com a capacidade de agrupar dados de pacientes em *clusters* com características similares, a clusterização pode ser utilizada na medicina para que, a partir de dados de saúde, identificar os vários tipos de agrupamentos de pacientes. As informações extraídas, que incluem grupos, distâncias, entre outras, podem ser interpretadas e, assim, facilitar o trabalho de profissionais do ramo da saúde. A fim de encontrar padrões em dados de pacientes e agrupá-los em clusters bem definidos para auxiliar na pesquisa sobre futuros diagnósticos e prevenir recidivas, o foco desse trabalho foi aplicar técnicas de clusterização em pacientes com diagnóstico de câncer em base de dados aberta. A subseção 2.1 apresenta uma breve descrição da base de dados escolhida, e a subseção 2.2 descreve o algoritmo utilizado.

[1] https://scikit-learn.org/stable/

2.1 Base de Dados

A base de dados escolhida para o projeto foi a base de dados de câncer da Fundação Oncocentro de São Paulo (FOSP)[2]. Ela é uma base aberta que acumula 1.066.999 dados estruturados de pacientes desde o ano 2000. A base da FOSP possui 99 *features*, isto é, 99 atributos com descrições dos pacientes. Os atributos incluem dados demográficos (ex: código do IBGE, estado e cidade), dados pessoais (ex: sexo, idade), informações sobre o tratamento (ex: combinação de tratamentos utilizados, como cirurgia, radioterapia e quimioterapia, diferença de dias entre consulta e diagnóstico e diferença entre diagnóstico e início do tratamento) e informações sobre o tumor (ex: lateralidade, código TNM, morfologia, topografia, metástases e recidiva).

Com uma vasta quantidade de informações, espera-se ser possível encontrar associações entre as *features* da base de dados e os casos de recidiva nos pacientes ao agrupá-los com o algoritmo de clusterização escolhido.

2.2 Algoritmo K-Prototypes

Em razão da grande quantidade de dados e da natureza majoritariamente nominal das features, o algoritmo utilizado na análise da base de dados foi o K-Prototypes [Huang 1998]. Como esse algoritmo que unifica o K-Means com o K-Modes, para entender seu funcionamento, é necessário explicar como o K-Means e como o K-Modes funcionam e, então, unir os conceitos no K-Prototypes.

O K-Means é um algoritmo de clusterização que, como critério de distância, utiliza da distância euclideana, definida como: $d(p_1, p_2) = \sqrt{\sum_{k=1}^{n}(p_{1k} - p_{2k})^2}$. Uma das características desse algoritmo é que quem o aplica precisa escolher um número *k* de clusters que serão formados. Ao iniciar, *k* pontos são escolhidos aleatoriamente, esses pontos serão os centroides dos clusters. Logo após, cada ponto do conjunto de dados é associado ao centroide em que a distância euclideana é a menor. Por último, a média dentre todos os pontos de cada cluster é calculada e o centroide passa a ser o ponto médio encontrado. O algoritmo repete esses passos até que o deslocamento dos centroides seja mínimo, ou um número máximo de iterações escolhido tenha sido atingido.

O K-Modes, assim como o K-Means, também recebe como entrada um número *k* de *clusters*, porém o critério de distância não é a euclideana, e pode ser aplicado para dados nominais. Ao iniciar, *k* pontos são escolhidos aleatoriamente dentro do conjunto para serem os centroides iniciais. Após, cada ponto é associado a um cluster, para isso, pega-se cada *feature* que constitui o ponto e calcula o número de diferenças para cada centroide, o ponto será designado para o cluster cuja centroide possuir menor diferença. Na sequência, um novo centroide é calculado juntando todos os pontos de cada cluster; para tanto, é calculada a moda de cada *feature*. Por fim, o algoritmo repete todos os passos até que todos os pontos fiquem estáveis em seus clusters, ou um número máximo de iterações escolhido tenha sido atingido.

Com todos os principais conceitos a respeito dos algoritmos envolvidos na formação do K-Prototypes, este é definido utilizando como métrica de dissimilaridade entre dois objetos a soma algébrica da distância euclideana para as

[2] http://www.fosp.saude.sp.gov.br/

features numéricas e a dissimilaridade do K-modes para as *features* nominais multiplicada de um fator *Y*, para assim balancear o peso entre *features* numéricas e categóricas.

3. Metodologia Utilizada

A metodologia implementada no projeto foi utilizar o algoritmo K-prototypes na base de dados de câncer da FOSP. Para isso, antes a base foi analisada e algumas *features* escolhidas, então o algoritmo foi aplicado e os resultados salvos para interpretação.

3.1 Escolha de *features* e pré-processamento

Primeiramente, para que o algoritmo tenha uma precisão mais efetiva, foi necessário separar a base de dados por tipo do câncer. Então, a quantidade de casos de cada topografia foi analisada, e o câncer de próstata, indicado por CID-O C619, foi escolhido para aplicação do K-Prototypes. A escolha se deu em função de conter grande quantidade de pacientes (*110.023* casos), bem como por simplificar a análise (ex: *feature* Sexo).

Em seguida, as 99 *features* foram analisadas e foram escolhidas aquelas que passavam a maior informação possível sobre os pacientes e poderiam influenciar com dados clínicos e demográficos na formação dos *clusters*. Dentre as *features* numéricas, a idade, a diferença de dias entre consulta e tratamento e entre diagnóstico e tratamento foram selecionadas. Já entre as *features* nominais, as escolhidas foram escolaridade do paciente, morfologia do tumor, base do diagnóstico, estádio clínico, categoria do atendimento (particular, SUS, convênio), perda ou não do segmento, código TNM, Gleason, tratamento no hospital, tratamento fora do hospital e presença de recidiva. As *features* selecionadas podem ser associadas, por exemplo, à recidiva de cncer por indicarem informações sobre a gravidade do tumor, os tratamentos realizados e como tais tratamentos foram aplicados.

Após separar a base de dados com as *features* desejadas, a biblioteca scikit-learn foi utilizada para fazer o pré-processamento das *features* numéricas, padronizando-as para que não haja um desbalanceamento e que uma *feature* não influencie mais que outras durante a clusterização. Por fim, todas os pacientes que possuíam informações faltando (NaN) foram removidos, resultando em uma nova base de dados com *99.908* casos, para análise do algoritmo.

3.2 Escolha do valor de *k* e algoritmo

Após separar as *features* numéricas e nominais e indicar para o algoritmo, o K-prototypes foi aplicado numa faixa de 1 a 5 clusters. Em seguida, a função custo de cada aplicação foi computado em um gráfico (Figura 1) junto com o número de clusters correspondentes. Para o melhor valor de *k*, foi escolhido aquele que a função custo decresce o mínimo possível.

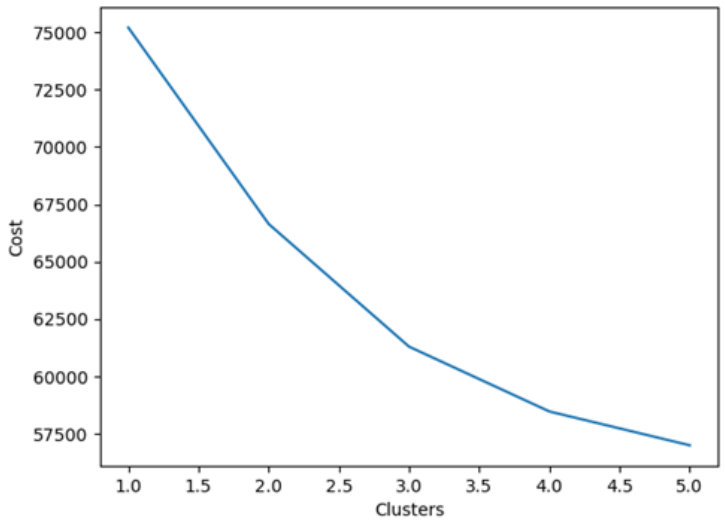

Figura 1. Custos versus número de Clusters

4. Resultados e interpretações preliminares

Analisando o gráfico apresentado na Figura 1, e a partir de testes, o valor de *k* foi selecionado como *5* e o algoritmo foi aplicado, gerando *5* centroides, e portanto, a Tabela 1 foi gerada com 5 *clusters*. Os centroides gerados podem ser visualizados na Tabela 1.

Tabela 1. Centroides gerados para os 5 *clusters*

CLUSTER	IDADE	TRATCON	DIAGTRA	ESCOLAI	MORFO	BASEDIAG	EC	CATEATEN	PERDASEG	T	N	M	G	PSA	GLEASON	TRATHOSP	TRATFAPOS	RECLOCAL
1	63	126	167	9	85503	3	IIB	2	0	2C	0	0	8	1	2	A	J	0
2	68	86	143	9	85503	3	IIA	2	0	1C	0	0	8	1	1	I	J	0
3	68	120	176	2	85503	3	III	2	0	3	0	0	8	8	8	I	J	0
4	64	90	104	2	85503	3	II	9	0	2C	0	0	8	8	8	A	J	0
5	70	67	109	9	81403	3	II	2	0	2	0	0	8	8	8	I	J	0

Para facilitar a interpretação, convém explicar o resultado das seguintes *features*:

- Escolaridade: 2 – Ensino Fundamental Incompleto e 9 – Ignorada
- Morfologia: 85503 - Carcinoma de células acinosas e 81403 - Adenocarcinoma, SOE (FOSP)
- Base Diagnóstico: 3 - Confirmação Microscópica
- Categoria de Atendimento: 2 – SUS e 9 – Sem informação
- Perda Segmento: 0 - Não e 1 – Sim
- T: 2 – Tumor confinado à próstata, 2C - Tumor que envolve ambos os lobos, 3 - Tumor que se estende através da cápsula prostática
- N: 0 - Ausência de metástase em linfonodo regional
- M: 0 - Ausência de metástase a distância

- G: 8 - Não se aplica
- PSA: 1 – Menor que 10 e 8 - Não se aplica

Analisando a quantidade de indivíduos por *cluster*, encontramos que o cluster 4 e 5 são os grupos mais frequentes, sendo de *26.521* e *25.490*, respectivamente. Ou seja, esses *clusters* concentram mais da metade dos casos de câncer de próstata. Dentre as recidivas, o cluster 4 apresenta *2137* casos e o cluster 5 apresenta *1725*, sendo essa quantidade somada *57%* dos casos de recidiva encontrados na base de dados estudada.

Após analisar as recidivas do cluster 4, podemos descrever como um grupo que se destaca por apresentar a maior quantidade de recidivas entre a faixa etária de *58* a *63* anos, possuir a morfologia 85503 (carcinoma de celulas acinosas) com *90%*, estádio clínico *II* e tratamento realizado no hospital sendo unicamente a cirurgia, portanto, parecido com o que o centroide do cluster 4 representa. Já o cluster *5* se destaca por possuir uma idade avançada, entre *64* e *75* anos, morfologia *81403* (adenocarcinoma SOE), estádio clínico II e tratamento no hospital como sendo de "outras combinações de tratamentos".

5. Conclusão

A partir dos resultados preliminares, entende-se que a clusterização pode ser utilizada como uma ferramenta computacional em saúde para agrupar pacientes em *clusters* bem definidos e, a partir dos grupos formados, discutir e interpretar os resultados de forma a apoiar a alguma conclusão sobre doenças comuns.

De acordo com a pesquisa, os clusters obtidos a partir do algoritmo K-prototypes resumem de forma inicial a base de dados ao apresentar 5 grupos distintos, porém com diversas características similares. Por exemplo, foram encontrados dois grupos que concentram uma quantidade significativa de pacientes e recidivas, podendo ser estudados à parte para uma análise mais minuciosa dos motivos. Também, um dos *clusters* encontrados (Cluster 2) possui uma quantidade de *445* recidivas (*6.5%* dos casos totais), podendo ser visto como um grupo com menor risco de recidiva se for comparado com os clusters *4* e *5* (*57%* dos casos).

Apesar dos resultados já obtidos, ao tratar dados derivados de um algoritmo, é preciso ter cuidado e analisar do ponto de vista médico para evitar conclusões incorretas. Portanto, o trabalho pode ser visto como uma motivação para a aplicação de técnicas computacionais mais avançadas ou estudos mais aprofundados no assunto, analisando do ponto de vista médico em conjunto com o aprendizado de máquina.

Para trabalhos futuros, pretende-se aplicar algoritmos mais complexos de clusterização, assim como estudar a fundo a estruturação da base de dados e tirar conclusões mais incisivas do assunto com participação de profissionais de saúde.

Referências

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

Landau, S., Leese, M., Stahl, D., & Everitt, B. S. (2011). Cluster analysis. John Wiley & Sons.

Panesar, A. (2019). Machine learning and AI for healthcare (pp. 1-73). Coventry, UK: Apress.