II SIMPÓSIO DE PRODUÇÃO CIENTÍFICA DA UNIFESSPA
A Pesquisa Científica e as Interações com a Realidade Amazônica
1 E 2 DE FEVEREIRO DE 2018

Clusterização de Dados Médicos: Definição do Perfil do Paciente com Câncer de Tireoide

Ariel Viana Silva[1] ; Regiane Leandro Pizon[1]; Haroldo Gomes Barroso[1]
[1]*Faculdade de Computação e Engenharia Elétrica, UNIFESSPA, 68507-590, Marabá-PA, Brasil*

Palavras-Chave: Câncer. Mineração de Dados. Clusterização.

1. INTRODUÇÃO

Com o advento dos Sistemas de Informação em Saúde (SIS), os hospitais e clínicas começaram a atuar em um ambiente informatizado, guardando Prontuários Eletrônicos e registros de exames laboratoriais em meios digitais.

As unidades hospitalares de Oncologia no Brasil, que atendem pacientes diagnosticados com Câncer, armazenam informações do prontuário eletrônico em uma base de dados nacional titulada RHC (Registros Hospitalares de Câncer). Esses registros fornecem informações sobre a doença, elementos demográficos e culturais do paciente, exames utilizados no diagnóstico, caracterização do tumor e tratamentos recebidos pelo paciente. O conjunto desses elementos (atributos) tornam possível a caracterização da doença.

Câncer é um termo genérico para um grande grupo de doenças que podem afetar qualquer parte do corpo. Uma característica do Câncer é o crescimento acelerado de células anormais além de seus limites habituais que podem então invadir outras partes do corpo [10].

A incidência elevada de casos de Câncer é um problema da saúde pública para países em desenvolvimento, no Brasil é a segunda maior causa de morte, com 190 mil óbitos por ano [5]. Em 2015 o Câncer foi uma das principais causas de morte no mundo, sendo responsável por 8,8 milhões de óbitos [10].

O Instituto Nacional de Câncer (INCA) estimou para o biênio 2016-2017 a ocorrência de cerca de 600 mil casos novos. Entre os 10 cânceres que mais aflige as mulheres está o da tireoide com uma estimativa de 5.870 casos novos [4].

A tireoide é uma glândula localizada na parte de trás do pescoço, abaixo das cordas vocais. Responsável pela produção de hormônios que ajustam o metabolismo, o surgimento de nódulos pode levar ao diagnóstico de Câncer. É a neoplasia maligna mais comum do sistema endócrino na faixa etária dos 30 aos 74 anos, possuindo prevalência três vezes maior no gênero feminino do que no masculino [11]. Deve ser dada atenção para o desenvolvimento biológico, irradiação prévia e história familiar de Câncer da tireoide [6].

Há um esforço constante da comunidade científica em auxiliar no controle do Câncer, pesquisas são desenvolvidas com apoio de técnicas de inteligência computacional para detecção de tumores malignos em exames de imagens [1]; para auxiliar no diagnóstico precoce da doença, com base em dados de resultados de exames [7]; e para determinar fatores de risco relacionados ao Câncer [9]. Esses avanços se tornaram possíveis devido ao crescimento do volume de dados médicos, possibilitando a aplicação de técnicas como a Mineração de Dados.

A Mineração de Dados, faz parte de um processo maior conhecido como KDD (*Knowledge Discovery in Databases*) – em português, Descoberta de Conhecimento em Bases de Dados –, que segundo Fayyad, Piatetsky-Shapiro e Smyth (1996)[3], é o processo não trivial de identificar em dados, padrões que sejam válidos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão.

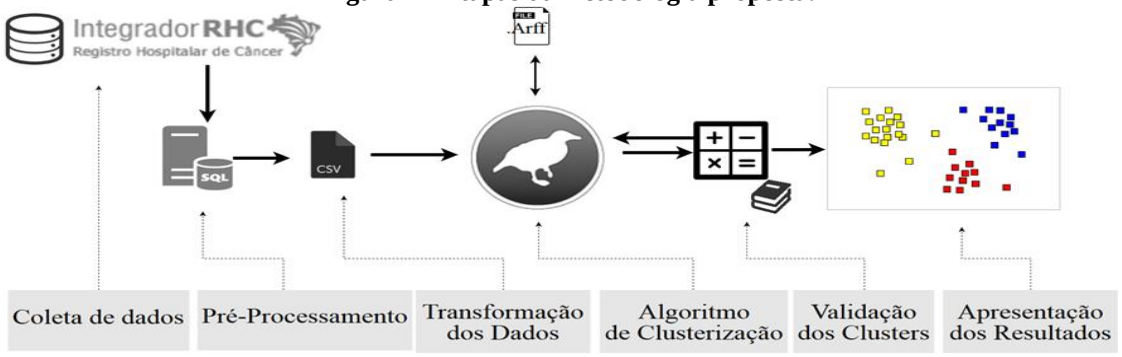
Uma forma de aplicar a Mineração de Dados é por meio da Clusterização, que busca formar diversos grupos conforme a similaridade dos atributos de um determinado universo, onde os atributos que partilham de características similares ficam em um mesmo grupo. No contexto do Câncer isso possibilita verificar as características similares que pacientes de um determinado Câncer possuem, facilitando a identificação do perfil.

Este trabalho visa analisar os dados da base RHC para mineração de dados, utilizando a tarefa de Clusterização, a fim de efetuar um estudo sobre o perfil dos pacientes com Câncer da tireoide. Levando em consideração os atributos de faixa etária, sexo, raça, consumo de álcool, consumo de tabaco e estadiamento do Câncer.

2. MATERIAL E MÉTODOS

Para que os objetivos estabelecidos fossem alcançados, fez-se necessário elaborar etapas para o desenvolvimento do trabalho (Figura 1).

Figura 1 - Etapas da metodologia proposta.



Fonte: Autor (2017)

Primeiramente os dados foram coletados do Integrador-RHC, sistema Web mantido pelo INCA (https://irhc.inca.gov.br). Em seguida, foi realizada a seleção de variáveis e pré-processamento dos dados em linguagem SQL (*Structured Query Language*), com o auxílio do PostgreSQL. Logo após, elaborou-se um arquivo CSV (*Comma-separated values*), adequação necessária para a utilização dos dados pelo software Weka (*Waikato Environment for Knowledge Analysis*), utilizado no processo de mineração de dados. Posto isto, aplicou-se o algoritmo de Clusterização EM (*Expectation Maximization*).

O EM [2] é um algoritmo iterativo usado para agrupar dados baseado em modelos de mistura. O EM atribui cada objeto ao cluster conforme parâmetros de probabilidade, para esse proposito o algoritmo alterna entre executar etapas de expectativa e maximização: primeiro, a etapa da expectativa (E) calcula os valores esperados para as probabilidades do cluster e, em segundo lugar, o passo de maximização (M) calcula os parâmetros de distribuição e a sua probabilidade conforme os dados. Itera até que os parâmetros sejam otimizados e alcancem um ponto fixo ou até

a função log-verossimilhança, que mede a qualidade do agrupamento, atinja seu máximo [8]. O algoritmo EM foi executado com a seguintes configurações:

Tabela 1 – Parâmetros utilizados para execução do algoritmo EM

Parâmetro	Descrição	Valor atribuído
Maxinterations	Número máximo de iterações.	500
numClusters	Número de Clusters a serem formados.	5
numKMeansRuns	Método de inicialização	50
minLogLikelihood	Melhoria mínima na verossimilhança para realizar outra iteração.	1.0E-6
minStdDev	Desvio padrão mínimo permitido.	1.0E-6

Fonte: Autor (2017).

3. RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados alcançados através da execução do algoritmo EM. Para o experimento foram consideradas 11.595 instâncias de Câncer da tireoide. O algoritmo convergiu em um tempo médio de 12,30 segundos, depois de 100 iterações, com log de verossimilhança igual a -10.17. A Tabela 2 apresenta os 5 clusters gerados.

Tabela 2 – Clusters de Câncer de tireoide.

Atributo	0	1	2	3	4
Instâncias	(0.16)	(0.25)	(0.29)	(0.19)	(0.11)
Sexo	F	F	F	F	F
Idade	45	45	45	46	52
Raça/Cor	Branca	Parda	Parda	Branca	Parda
Histórico familiar	S	S	N	N	N
Alcoolismo	Nunca	Nunca	Nunca	Nunca	Sim
Tabagismo	Nunca	Nunca	Nunca	Nunca	Sim
TNM	100	100	100	100	100

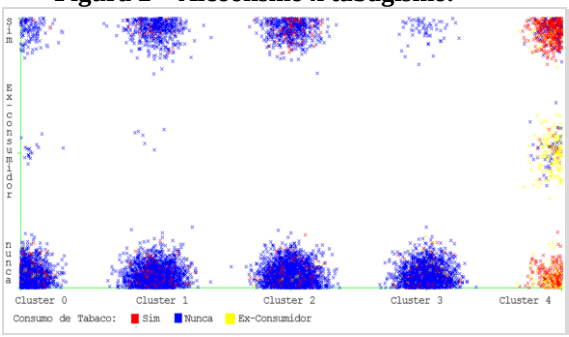
Fonte: Autor (2017)

O Cluster 2 (29% das ocorrências), caracteriza a maior parte dos pacientes, que são do sexo feminino, com idade média de 45 anos, raça/cor parda, sem histórico prévio de Câncer na família, consumo de bebidas alcoólicas ou tabaco. Esse Câncer é descoberto no estágio 100, conforme o sistema de Classificação dos Tumores Malignos (TNM), onde o primeiro item T, representa a extensão do tumor, com valores de 1-4; N, se refere ausência ou presença de metástases nos linfonodos, com valores de 1-3; M, se refere à ausência ou presença de metástase à distância, com valores de 0 ou 1.

O EM agrupou os indivíduos fumantes no menor cluster. Na Figura 2 é observado a presença de pessoas consumidoras de bebida alcoólica e tabaco no *cluster* 4, que representa 11% da amostra.

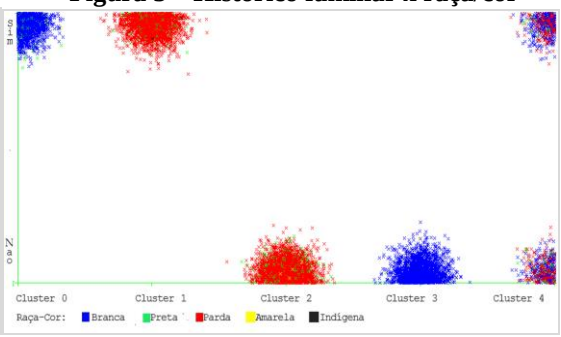
Pessoas pardas foram unidas em 3 grupos (1, 2 e 4), que juntos correspondem a 65% das instâncias, conforme a Figura 3, que também apresenta a ocorrência do histórico de câncer na família, presente principalmente nos *clusters* 0 e 1.

Figura 2 – Alcoolismo x tabagismo.



Fonte: Autor(2017).

Figura 3 – Histórico familiar x raça/cor



Fonte: Autor(2017).

4. CONCLUSÃO

Neste trabalho aplicou-se a técnica de Clusterização de dados para identificar em pacientes com os Cânceres da tireoide, a presença das variáveis: sexo, idade, histórico familiar, tabagismo, alcoolismo e TNM. Conforme os resultados, constata-se que o Câncer da tireoide incide principalmente pessoas do sexo feminino, com médias de idade entre 45 e 52 anos, com raça/cor parda, seguida pela branca. O consumo de álcool e tabaco é um hábito da minoria dos indivíduos com esse Câncer e o seu tratamento geralmente é iniciado no estágio 100, conforme a classificação TNM.

REFERÊNCIAS

CHATTARAJ, A.; DAS, A.. Mammographie image segmentation using kernel based FCM clustering approach. In: *Computer, Electrical & Communication Engineering* (ICCECE), 2016 International Conference on. IEEE, 2016. p. 1-6.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. *Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society*. Series B (methodological), p. 1-38, 1977.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.

INCA. a. Estimativa 2016: incidência de Câncer no Brasil. Instituto Nacional de Câncer José Alencar Gomes da Silva – Rio de Janeiro: INCA, 2015. 122p. ISBN 978-85-7318-283-5.

INCA. b. Instituto Nacional do Câncer. INCA estima quase 600 mil casos novos de Câncer para 2016. Rio de Janeiro: INCA, 2015. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/agencianoticias/site/home/noticias/2015/inca_estima_quase_600_mil_casos_novos_de_cancer_em_2016>. Acesso em: mar. 2017.

INCA. Câncer da Tireóide: *Thyroid cancer*. Revista Brasileira de Cancerologia, v. 48, n. 2, p. 181-185, 2002.

JOSHI, J.; DOSHI, R.; PATEL, J. Diagnosis of breast cancer using clustering data mining approach. *International Journal of Computer Applications*, (0975– 8887) v. 101, n. 10, 2014.

MARKOV, Z.; LAROSE, D. T. *Data mining the Web: uncovering patterns in Web content, structure, and usage*. John Wiley & Sons, 2007.

NAHAR, J. et al. *Brain Cancer Diagnosis-Association Rule Based Computational Intelligence Approach*. In: *Computer and Information Technology* (CIT), 2016 IEEE International Conference on. IEEE, 2016. p. 89-95.

ORGANIZAÇÃO MUNDIAL DA SAÚDE - OMS. *Cancer*. fev. 2017. Disponível em: <http://www.who.int/mediacentre/factsheets/fs297/en/>. Acesso em: mar. 2017.

VIANNA, D. M. et al. *The histological rarity of thyroid cancer. Brazilian journal of otorhinolaryngology*, v. 78, n. 4, p. 48-51, 2012.