Biomatemática ***28*** *(2018), 15–24* ***ISSN 1679-365X***
Uma Publicação do Grupo de Biomatemática IMECC – UNICAMP

Modelo de classificação para risco de metástase em pacientes com câncer de rim usando mineração de dados

Cristina Sacilotto[1],
DMA, IMECC – Unicamp, 13.083-859, Campinas/SP.

Diego F. Gomes[2],
DEMAT, IFMA, 65.950-000, Barra do Corda/MA.

Stanley R. de M. Oliveira[3],
Embrapa, FEAGRI – Unicamp, 13083-886, Campinas/SP.

Ubirajara Ferreira[4],
FCM – Unicamp, 13083-887, Campinas/SP.

Graciele P. Silveira[5],
DFQM – UFSCar, 18.052-780, Sorocaba/SP.

Rodney C. Bassanezi[6],
DMA, IMECC – Unicamp, 13.083-859, Campinas/SP.

Laércio L. Vendite[7],
DMA, IMECC – Unicamp, 13.083-859, Campinas/SP.

Resumo. Os tumores renais malignos têm importante significado clínico e os mais frequentes são os carcinomas de células renais (CCR). Analisou-se a relação entre determinados fatores prognósticos (principalmente a graduação de Fuhrman) e a sobrevida desses pacientes, tendo em vista que constatou-se,

[1]cristinasacilotto@gmail.com
[2]diego.gomes@ifma.edu.br
[3]stanley.oliveira@embrapa.br
[4]ubirafer@uol.com.br
[5]graciele@ufscar.br
[6]rodney@ime.unicamp.br
[7]vendite@ime.unicamp.br

em casos reais, pacientes com graus de Fuhrman baixos (considerados satisfatórios) e apresentaram um prognóstico desfavorável. As variáveis foram escolhidas com base em referências na área e em consulta a especialistas. Os dados utilizados foram obtidos do Hospital das Clínicas da Unicamp – HC. Foram aplicadas técnicas de mineração de dados através dos algoritmos Apriori com classe indexada e C4.5 a fim de obter um modelo para predizer o risco de desenvolvimento de metástase pelos pacientes que apresentam a doença com o subtipo convencional (células claras). Foi feita uma análise de correlação atributo-atributo para verificar possíveis redundâncias nos dados. Classificar o risco de metástase é um desafio, já que o conjunto de dados tem poucos exemplos dessa classe (classe positiva ou minoritária). Por essa razão, técnicas para balanceamento de classes foram utilizadas para garantir que os algoritmos inteligentes aprendessem os padrões da classe minoritária. Os resultados obtidos foram satisfatórios, pois foi possível ranquear os atributos e identificar aqueles com maior ganho de informação. Concluímos de maneira investigativa que o atributo Tumor (tamanho) possui alta correlação com Estadiamento, o que constatou a retirada desse atributo. Além disso, a análise dos dados contribuiu para determinar que a graduação de Fuhrman é o segundo atributo em ordem de importância para predizer o risco de metástase.

***Palavras-chave**: Câncer de rim; Data mining; Modelagem; Biomatemática.*

1. Introdução

O câncer é a 2ª principal causa de morte em todo o mundo em comparativo a outras doenças registradas pela Organização Mundial de Saúde (OMS). Em 2012, foi registrado 8,2 milhões de mortes relacionadas ao câncer em todo o mundo, uma estimativa de 13% de todas as mortes mundiais (World Health Organization – WHO, 2016). Em particular, as neoplasias renais malignas têm importante significado clínico e são responsáveis por cerca de 2% dos tumores malignos em humanos sendo mais frequentes os carcinomas de células renais (CCR). Entre os subtipos geneticamente diferentes, destaca-se o convencional (células claras), que é o subtipo mais comum e objeto de estudo desse trabalho.

Os fatores prognósticos relacionados ao câncer de rim apresentam várias controvérsias e imprecisões, pois estão ligados à biologia do tumor. Um dos fatores com considerável valor prognóstico é o Sistema de Graduação de Fuhrman (Fuhrman et al., 1982) o qual classifica o padrão nuclear celular da neoplasia

em quatro graus.

Além dessa série de parâmetros considerados fatores prognósticos importantes, acredita-se que o estádio da doença tem valor prognóstico e pode influenciar a sobrevida dos pacientes. Recomenda-se a classificação UICC TNM 2009 (Tumour Node Metastasis) para o estadiamento do CCR, que está descrito em (Ljungberg, 2009).

Uma das alternativas promissoras para classificar pacientes com CCR é utilizar técnicas de mineração de dados. Mineração de dados é uma área de pesquisa multidisciplinar (herda métodos e técnicas da estatística, matemática, inteligência artificial, etc) e tem como objetivo central extrair conhecimento (padrões e tendências) de grandes volumes de dados. Essas técnicas têm sido fortemente utilizadas para auxiliar decisões estratégicas em diversas áreas do conhecimento (Han et al., 2011).

Usaremos as técnicas de mineração de dados para extrair informação com base em um Banco de Dados (BD) pré-processados, obtido da tese de mestrado da autora Sacilotto (Sacilotto, 2017), para gerar conhecimento sobre o risco de metástase em pacientes com câncer de rim.

2. Objetivos

- Analisar a relação entre a graduação de Fuhrman e o prognóstico de pacientes com tumor renal;
- Analisar a relação entre os atributos a serem considerados como forma de gerar regras para diagnosticar o risco de metástase.

3. Metodologia

Objetivando obter o risco de metástase como meta (ou variável resposta), modificamos os valores da variável-meta em dois resultados, baixo para valores iguais a zero ou três os quais pertencem aos grupos de pessoas que não tiveram metástase em nenhuma etapa do tratamento, e alto para valores iguais a um ou dois os quais tiveram metástase em alguma etapa do tratamento.

Os atributos foram discretizados tornando-se todos categóricos, ou seja, diâmetro do Tumor foi renomeado como P (0–4 cm), M (4–7 cm), MG (7–10 cm) e G (>10 cm); Necrose foi renomeada como zero (sem necrose) e um

(com necrose); graduação de Fuhrman pelo seu tipo: um, dois, três e quatro; e Estadiamento já era nominal: T1a, T1b, T2a, T2b, T3a, T3b e T4.

A princípio tínhamos um total de 132 instâncias, mas ao analisar os resultados um a um notamos algumas inconsistências que foram corrigidas com a ajuda de um especialista o qual indicou a retirada de uma instância como possível outlier*.

Para tratamento dos dados coletados usamos o software de domínio público, *Waikato Environment for Analysis* (WEKA) (University of Waikato, 1993), da Universidade de Nova Zelândia.

Como passo inicial na busca de nossos objetivos, aplicamos o algoritmo Apriori (Liu et al., 1998) com o atributo-meta indexidado, suporte igual a $0,1$ e confiança igual a $0,9$. Observamos que o algoritmo gerou 25 regras onde somente a classe majoritária (risco de metástase baixo) foi classificada, resultado já esperado devido o desbalanceamento das classes observado no histograma gerado pelo WEKA (92 instâncias classificadas como risco baixo e 39 classificadas como risco alto). Usamos, também, o algoritmo C4.5 (Quinlan, 1993) (J48 no WEKA) com poda para gerar regras que classifiquem bem este risco tendo como cojunto de teste a validação cruzada com 10 folds (Kohavi, 1995), visto que nosso banco de dados é pequeno. Os dois algoritmos mencionados foram escolhidos para realizarmos nossa análise devido sua inteligibilidade.

Em busca de melhorar os resultados iniciais obtidos, investigamos a existência de atributos correlacionados usando o Teste do Qui-quadrado (Ugoni e Walker, 1995), pelo fato de nossos atributos serem nominais, com a intenção de manter somente aqueles correlacionados com o atributo-meta.

Após a análise de dimensão usamos técnicas de amostragem para tratarmos do desbalanceamento das classes (Batista et al., 2004): Random Sampling (Resample do WEKA), NCL e SMOTE. Estes modelos foram comparados pela acurácia e pela *Area Under the Curve* (AUC) (Prati et al., 2008) da classe positiva gerados pelo algoritmo C4.5 com validação cruzada de 10 folds e, então, foi escolhido aquele que apresentou os melhores resultados.

Na próxima seção faremos a análise de algumas etapas aqui mencionadas.

*São objetos com características diferentes da maioria dos outros objetos em um conjunto de dados.

4. Resultados

Apesar de nosso banco de dados ter quatro atributos, fora o atributo-meta, fizemos o Teste do Qui-quadrado em dois a dois atributos apresentados na Tabela 1.

Tabela 1: Valores do Teste do Qui-quadrado de dois a dois atributos.

Atributos	**Qui-quadrado** (χ^2)
Tumor – Necrose	$11,475$
Tumor – Fuhrman	$32,0396$
Tumor – Estadiamento	$260,92$
Necrose – Fuhrman	$24,527$
Necrose – Estadiamento	$15,194$
Fuhrman – Estadiamento	$40,114$

Notamos que os atributos Tumor e Estatiamento possuem um alto grau de correlação. Temos que o Estadiamento é calculado com base no tamanho do tumor. Fizemos testes usando o algoritmo C4.5 e notamos que a acurácia do modelo sem a retirada do atributo foi igual a $87,7863\%$ e com a retirada do atributo a acurácia foi igual a $86,2595\%$. Fizemos o teste com o inverso, retirar o atribito Estadiamento, e a acurácia foi igual a $87,0229\%$. Para sairmos deste impasse usamos o Teste do Qui-quadrado[†] para ranquear os atributos em relação ao atributo-meta e obtemos os seguintes resultados apresentados na Tabela 2 .

Tabela 2: Ranque do Teste do Qui-quadrado dos atributos em relação ao atributo-meta.

Ranque	**Atributo**
$57,2888$	Estadiamento
$46,0492$	Tumor
$34,8068$	Fuhrman
$17,9208$	Necrose

Portanto, notamos que o Estadiamento fala mais sobre o Risco de Metástase que o Tumor.

[†]O InfoGain e o GainRatio geraram o mesmo ranque com diferentes valores.

Usamos técnicas de amostragem para corrigir o desbalanceamento das classes em cima de três tipos de BD: sem a retirada de atributos, com a retirada do atributo Tumor e com a retirada do atributo Estadiamento. E logo após, fizemos uma análise observando a acurácia de cada modelo criado usando o algoritmo C4.5 com validação cruzada (10 folds) cujos resultados são vistos na Tabela 3.

Tabela 3: Acurácia do algoritmo C4.5 aplicado ao BD com e sem a retirada de atributos com diferentes técnicas de amostragem para classes desbalanceadas.

Técnicas de Amostragem	**Banco de Dados**		
	Sem Retirada	**Sem Estadiamento**	**Sem Tumor**
Resample	92,3664%	93,1298%	93,8931%
NCL	96,6667%	92,6316%	99,1071%
SMOTE	92,9412%	87,0588%	89,4118%
NCL-SMOTE	98,7421%	94,0299%	100%
SMOTE-NCL	98,7500%	94,0299%	100%
Sem amostragem	87,7863%	87,0229%	86,2595%

Classificar o risco de metástase como alto em pacientes com câncer de rim é mais difícil que classificá-lo como baixo, portanto vemos que a classe alto é a classe positiva. Com efeito, observar somente a acurácia dos resultados pode levar-nos a erros gravíssimos, i.e., modelos que classifiquem somente a classe majoritária podem possuir valores altos de acurácia. Portanto, com base nos modelos analisados acima vamos analisar também a *Area Under the Curve* (AUC) da classe positiva, i.e., a área abaixo da curva ROC (*Receiver Operating Characteristics*) da classe positiva cujos resultados são vistos na Tabela 4.

Tabela 4: AUC da classe positiva do algoritmo C4.5 aplicado ao BD com e sem a retirada de atributos com diferentes técnicas de amostragem para classes desbalanceadas.

Técnicas de Amostragem	**Banco de Dados**		
	Sem Retirada	**Sem Estadiamento**	**Sem Tumor**
Resample	97, 93%	94, 42%	99, 29%
NCL	98, 08%	96, 89%	100%
SMOTE	93, 98%	91, 28%	93, 95%
NCL-SMOTE	99, 93%	96, 42%	100%
SMOTE-NCL	99, 91%	96, 42%	100%
Sem amostragem	91, 30%	89, 99%	88, 60%

Com base nas Tabelas 3 e 4 temos que o banco de dados sem o atributo Tumor com as técnicas somadas NCL-SMOTE ou SOMTE-NCL geram resultados muito bons. Temos também que as árvores de decisão, vista na Figura 1, geradas pelas duas técnicas (diferindo a ordem de aplicação) são as mesmas, i.e., mesmas regras para classificação.

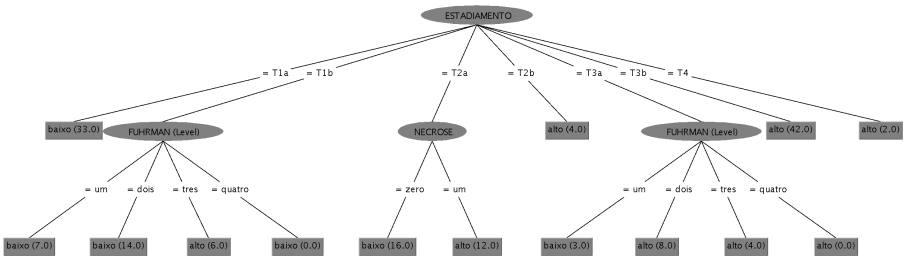

Figura 1: Árvore de decisão gerada pelo algoritmo C4.5 após aplicação do NCL-SMOTE ou SMOTE-NCL.

Notamos alguns problemas nas regras criadas pela árvore acima. Olhando para a regra, Estatiamento = T1b e Fuhrman = quatro classificou-se o risco de Metástase como baixo o que, provavelmente, deveria ser classificado como alto. Isso acontece pelo fato de não termos nenhum exemplo que se encaixasse nesta regra, portanto o algoritmo classificou como baixo pois foi a moda entre as classificações do ramo Estadiamento = T1b e Fuhrman. O mesmo aconteceu com a regra Estatiamento = T3a e Fuhrman = quatro que classificou como alto risco de metástase que é a moda do ramo Estatiamento = T3a e

Fuhrman. Este problema poderia ser melhorado com um aumento de exemplos (instâncias) para treinamento, mas como o nível de registros desta doença é pequeno não obtivemos essa generalização na regra gerada.

Usamos também o algoritmo Apriori com classe indexada, suporte igual a $0,1$ e confiança igual a $0,9$ sobre o banco de dados sem o atributo Tumor e com a aplicação dos filtros NCL-SMOTE e obtivemos 12 regras das quais 6 classificam a classe positiva.

```
Best rules found:

 1.  ESTADIAMENTO=T3b 42 ==>  METASTASE (Class)=alto 42    conf:(1)
 2.  ESTADIAMENTO=T1a 33 ==>  METASTASE (Class)=baixo 33    conf:(1)
 3.  FUHRMAN (Level)=um 31 ==>  METASTASE (Class)=baixo 31    conf:(1)
 4.  NECROSE=zero  ESTADIAMENTO=T1a 31 ==>  METASTASE (Class)=baixo 31    conf:(1)
 5.  NECROSE=zero  FUHRMAN (Level)=um 29 ==>  METASTASE (Class)=baixo 29    conf:(1)
 6.  NECROSE=zero  ESTADIAMENTO=T3b 28 ==>  METASTASE (Class)=alto 28    conf:(1)
 7.  FUHRMAN (Level)=tres  ESTADIAMENTO=T3b 26 ==>  METASTASE (Class)=alto 26    conf:(1)
 8.  NECROSE=um  FUHRMAN (Level)=dois 18 ==>  METASTASE (Class)=alto 18    conf:(1)
 9.  NECROSE=zero  FUHRMAN (Level)=tres  ESTADIAMENTO=T3b 18 ==>  METASTASE (Class)=alto 18    conf:(1)
10.  FUHRMAN (Level)=um  ESTADIAMENTO=T1a 17 ==>  METASTASE (Class)=baixo 17    conf:(1)
11.  NECROSE=zero  ESTADIAMENTO=T2a 16 ==>  METASTASE (Class)=baixo 16    conf:(1)
12.  NECROSE=um 36 ==>  METASTASE (Class)=alto 34    conf:(0.94)
```

Figura 2: Regras geradas pelo algoritmo Apriori com classe indexada, suporte igual a $0,1$ e confiança igual a $0,9$ sobre o banco de dados sem o atributo Tumor e com a aplicação dos filtros NCL-SMOTE.

A inviabilidade deste procedimento para nosso conjunto de dados é dada pelo fato da criação de regras redundantes, por exemplo, as regras 1, 6, 7, e 9, poderiam ser resumidas a regra 1 diminuindo ainda mais o poder de classificação, e por termos um banco de dados pequeno para este algoritmo.

5. Conclusões

Pela análise dos atributos sobre estes dados coletados de câncer de rim, concluímos que a graduação de Furhman foi o segundo atributo mais informativo quando eliminado o atributo diâmetro do Tumor. Temos que o Estadiamento é o atributo com maior ganho de informação (visto na Tabela 2) sobre o risco de metástase (atributo-meta) em pacientes com cancêr de rim com subtipo convencional, por este motivo, o Estadiamento é a raiz da ávore de decisão (Figura 1).

A princípio, pela pouca quantidade de atributos, achavamos que não seria necessária a eliminação de algum desses. A mineração dos dados nos

motrou claramente que a melhor hipótese construída é obtida quando eliminada o atributo Tumor.

A hipótese construída é uma boa proposta para médicos que acompanham pacientes com câncer de rim do subtipo convencional avaliarem o risco de metástase.

Agradecimentos

O primeiro e segundo autor agradecem a todos os professores coautores pelo ensino e idealização deste trabalho. O primeiro autor agradece a CAPES pela bolsa de doutorado e o segundo autor agradece ao Instituto Federal do Maranhão – IFMA.

Referências

Batista, G., Prati, R., e Monard, M. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1):20–29.

Fuhrman, S. A., Lasky, L. C., e Limas, C. (1982). Prognostic significance of morphologic parameters in renal cell carcinoma. *The American Journal of Surgical Pathology*, 6(7):655–663.

Han, J., Kamber, M., e Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In International joint Conference on Artificial Intelligence*, 14:1137–1145.

Liu, B., Hsu, W., e Ma, Y. M. (1998). Integrating classification and association rule mining. *Proceedings of the fourth international onference on knowledge discovery and data mining (KDD-98)*, páginas 80–86.

Ljungberg, B. (2009). Orientações sobre carcinoma das células renais. URL: http://www.uroweb.org. Acesso em: 15/09/2016.

Prati, R. C., Batista, G. E. A. P. A., e Monard, M. C. (2008). Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, 6(2):215–222.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Sacilotto, C. (2017). Uso da teoria de conjuntos fuzzy para análise prognóstica do câncer de rim. Dissertação de Mestrado, IMECC–UNICAMP, Campinas/SP.

Ugoni, A. e Walker, B. F. (1995). The chi square test: An introduction. *COMSIG review*, 4(3):61–64.

University of Waikato (1993). Weka the University of Waikato. URL: https://www.cs.waikato.ac.nz/ml/weka/downloading.html Acesso em: 20/09/2017.

World Health Organization – WHO (2016). What is cancer? URL: http://www.who.int/cancer/en/. Acesso em: 02/06/2017.