AVALIAÇÃO DO MÉTODO NAIVE BAYES NA CLASSIFICAÇÃO DE AMOSTRAS DA BASE DE DADOS ORIGINAL DO CÂNCER DE MAMA DE WISCONSIN

Ingrid Bergmam do Nascimento Silva[1]
Natasha Seleidy Ramos de Medeiros[2]
Ingrid Rafaella dos Santos Melo[3]
Larissa Duarte de Britto Lira[4]
Ronei Marcos de Moraes[5]

RESUMO

A análise de variáveis e a escolha de critérios com a finalidade de tomar decisões é um processo complexo, pois geralmente os dados apresentam-se de forma imprecisa e incompleta, com diversos agentes de decisão e objetivos que dificultam a tomada de decisão. O algoritmo naive bayes apresenta-se como um ótimo tipo de algoritmo para grandes valores de dados, também apresenta agilidade na execução, tudo isto quando comparado a outros algoritmos de classificação. Objetivo: neste estudo, foi avaliado o desempenho do método de decisão naive bayes na classificação da base de dados original de wisconsin. Metodologia: trata-se de um estudo descritivo e exploratório, o qual utilizou o método naive bayes através do weka com a utilização das técnicas de c*ross-validation, percentage split e random seed* para separação dos dados de treino do teste. resultados e discussão: diante da realização dos testes o que resultou no melhor kappa foi o *percentage split* com slip 85% e *random seed* 3 sendo coeficiente kappa=95,77%, já o menor kappa foi obtido através do teste *percentage split* com split 75% e *random seed* 1, sendo coeficiente kappa=89,09% em todos os testes realizados foi obtido um coeficiente kappa quase perfeito. Considerações finais: o algoritmo naive bayes apresenta-se como um ótimo classificador de dados, dessa forma, visto que a base de dados analisada é da área de saúde, sugere-se que essa técnica seja empregada na área da ciência da saúde a fim de utilizar como classificador para auxiliar no processo de tomada de decisão em saúde.

Palavras-Chave: Redes Bayesianas. Naive Bayes. Base de dados Wisconsin.

INTRODUÇÃO

A análise de variáveis e a escolha de critérios com a finalidade de tomar decisões é um processo complexo, pois geralmente os dados apresentam-se de forma imprecisa e incompleta, com diversos agentes de decisão e objetivos o que vem a dificultar a tomada de decisão. Diante disso é necessário avaliar as alternativas disponíveis para escolher a mais viável, podendo esta ser influenciada pela qualidade das informações disponíveis, pelo tempo e pelos recursos (MARTINS, 2014).

[1]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, ingridgba2006@hotmail.com;
[2]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, natashaseleidy@gmail.com;
[3]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, ingridmeello@gmail.com;
[4]Mestranda em Modelos de Decisão e Saúde da Universidade Federal da Paraíba- UFPB, larissadblira@hotmail.com;
[5]Professor da graduação e Pós-graduação no Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba- UFPB, ronei@de.ufpb.br.

Os algoritmos Bayesianos, denominados Naive Bayes (NB) é um classificador probabilístico ingênuo (naive) que surgiu a partir do teorema de Bayes (Thomas Bayes), nele os atributos apresentam-se independentes entre si, sendo a informação de cada evento não informativa sobre nenhum outro evento. Este algaritmo é considerado como um dos mais precisos e eficientes pois dado um determinado conjunto, onde este é composto por grupos divididos por atributos e valores, diante disto é possível saber a qual grupo pertence uma nova instância (CARVALHO; CIANN, 2013).

O algoritmo Naive Bayes apresenta-se como um ótimo tipo de algoritmo para grandes valores de dados, como também apresenta agilidade na execução, tudo isto quando comparado a outros algoritmos de classificação. Um classificador do tipo Naive Bayes admite que a presença de uma característica particular em uma classe não está relacionada com a presença de qualquer outro recurso (SÁ et al, 2018).

O classificador Naive bayes geralmente apresenta grande desempenho, isso deve-se ao fato de que este algoritmo considera cada atributo de forma independe, o que possibilita ser usado para trabalhar com informações incompletas e imprecisas (FACELI, 2011).

O Weka é classificado como um sistema de referência na história das comunidades de pesquisa de mineração de dados e de aprendizado de máquina, sendo ele o único kit de ferramentas que ganhou ampla adoção e sobreviveu por um longo período de tempo. O Weka ou Woodhen (Gallirallus australis) é uma ave endêmica da Nova Zelândia. O mesmo fornece muitos algoritmos diferentes para dados mineração e aprendizado de máquina. Weka é open source e está disponível para acesso gratuitamente além de ser independente de plataforma (SAPN; M, 2010).

Sistemas Inteligentes apresentam-se como uma ferramenta importante para auxiliar gestores no processo de tomada de decisão. Técnicas baseadas em aprendizado de máquina conseguem prever alguns eventos com certa precisão (GOLDSCHMIDT et al., 2015).

As Redes Bayesianas representam através do relacionamento de proposições ou variáveis que envolvem incerteza ou imprecisão, portanto o uso adequado de independências condicionais permite computação rápida e eficiente de probabilidades atualizadas para estados de variáveis não observadas (HOJSGAARD et al., 2012). Entretanto o classificador Naive Bayes consiste em um classificador probabilístico, apresenta robustez e boa eficiência tendo seus atributos independentes entre si (CARVALHO; CIANN, 2013).

As Redes Bayesianas apresentam-se em forma de grafo acíclico direcionado, onde cada um dos nós representa uma variável do domínio, ou seja, um evento específico com estados mutuamente exclusivos que pode ocorrer no domínio e que tem importância no contexto da

modelagem. As arestas representam as causas e/ou influências que um nó tem sobre outros nós. Cada nó possui uma tabela de probabilidades anexada onde este indica as probabilidades do evento em questão acontecer. Caso o evento não seja dependente de nenhum outro evento (ou seja, se não existir nenhuma relação ou aresta direcionada de algum outro nó para o nó que representa o evento), assim a tabela de probabilidades representa as probabilidades a priori ou marginais para o evento e é denominada de Tabela de Probabilidades Prévia. Porém se houver uma ou mais relações causais (arestas) de outros eventos (nós) para um evento (nó), deste modo a tabela de probabilidades irá definir as probabilidade condicionais associadas ao evento, sendo chamada de Tabela de Probabilidade Condicional do nó. Os nós com arcos dirigidos para o nó que representa o evento, são denominados de nós pais (VIER et al., 2015).

Raschka (2014) define Naive Bayes como sendo uma técnica de classificação de dados cuja fundação de seu funcionamento tem embasamento no teorema de Bayes.

Hazzan (1995) apresenta P(A|B) como a probabilidade condicional do evento A dado que B já ocorreu. No escopo da classificação de dados do Naive Bayes, não estamos lidando com dois eventos, mas sim um conjunto de classes as quais irão categorizá-los e com um conjunto de dados. Segundo Raschka (2014), para dado vetor de dados X com seus elementos x_i onde i corresponde a cada elemento do vetor e i = {1,2,..., n}, e ω_j a notação de classe tal que j = {1,2,..., m}, assim $P(\omega_j|X)$ é calculado por:

$$P(\omega_j|X) = \frac{P(x_i|\omega_j) * P(\omega_j)}{P(x_i)} \qquad (1)$$

Para Raizada e Lee (2013), o Naive Bayes baseia-se na teoria de que o valor de um atributo não influencia o valor de outros atributos, isto é, os mesmos sendo classificados são independentes entre si. É um algoritmo conhecido como ingênuo (naive) pois admite que os atributos são independentes, porém mesmo sendo tido como "ingênuo" é um dos classificadores mais utilizados para categorizar textos, devido a sua rotulação de novas amostras e de seu processamento (PUURULA, 2012), ademais o método é capaz de obter resultados tão bons quanto ou até melhores que modelos mais complicados que ele (RASCHKA, 2014). Com isto, levando em consideração a suposição de independência entre atributos, o cálculo do valor de P(x | j) se torna simplificado:

$$P(X|\omega_j) = P(x_1|\omega_j) * P(x_2|\omega_j) * \ldots * P(x_d|\omega_j) = \prod_{k=1}^{d} P(x_d|\omega_j) \qquad (2)$$

Dado que o valor de $P(x)$ é constante para uma dada instância do método, tem-se que a classificação pode ser realizada sem levar em consideração o seu valor. Moraes e Machado (2012) determinam a regra de classificação de uma Naive Bayes, apresentada a seguir com devidos ajustes à notação utilizada, como:

$$X \in \omega_i \; se \; P(\omega_i | \, X) > \; P(\omega_j | \, X), para \; todo \; i \neq j$$

A rede probabilística fuzzy naïve-bayes (PFNB) foi implementada baseada no modelo de sistema fuzzy probabilístico proposto por Kaymak e Berg (2013) e nos conceitos de inferência bayesiana. Esta rede é capaz de modelar ao mesmo tempo a imprecisão linguística e a incerteza estocástica, onde a partir da interpretação produz conhecimento com base no conjunto de dados.

Diante do exposto o estudo tem como objetivo avaliar o método de decisão Naive Bayes (NB), e testar se o mesmo se adequa a base de dados original do câncer de mama de Wisconsin.

METODOLOGIA

Trata-se do estudo da aplicação do método Naive Bayes sobre os dados fornecidos pelo banco de dados *Wisconsin* (original), disponível em um repositório internacional de banco de dados que pode ser acessado por meio do link: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). Este banco de dados possui 699 instâncias e 10 atributos e 2 classes.

Os dados são acerca do tema câncer de mama, e tem como criador o médico Dr. William H. da Universidade de Wisconsin Hospitais Madison em Wisconsin nos EUA. O banco de dados, portanto, reflete um agrupamento cronológico dos dados.

O software utilizado para aplicação da NBG sobre o banco de dados foi o programa WEKA (Waikato Environment for Knowledge Analysis), versão 3.8, desenvolvido na Universidade de Waikato na Nova Zelândia. Este possui vários algoritmos de modelos de decisão (J48, Id3, BFTree, User Classifier, Rede Naive Bayes, entre outros) e regressão (WEKA, 2017).

O WEKA possui vários algoritmos de modelos de decisão (J48, Id3, BFTree, User Classifier, Rede Naive Bayes, entre outros) e regressão (SILVA, 2004; WEKA, 2016). Ademais, faz relatórios com dados analíticos e estatísticos, pois possui uma interface gráfica amigável e uma pequena limitação na sua escalabilidade, já que suas versões atuais limitam o volume de dados a ser manipulado à dimensão de memória principal.

Segundo Bouckaert *et al.* (2016) a primeira reordena os dados aleatoriamente e depois os separa em *n* conjuntos de tamanhos iguais (o valor de *n* é determinado pelo usuário), destes *n* conjuntos, *n-1* são designados para treinamento e o outro para teste. A segunda técnica, que é a *percentage split*, para uma dada porcentagem definida pelo usuário, designa a porcentagem informada de dados para treinamento e o restante para testes, sendo que os dados são previamente reordenados aleatoriamente. O último parâmetro a ser utilizado para realizar os testes foi o *random seed*, esse parâmetro é utilizado para a reordenação aleatória dos dados e precisa ser diferente para cada teste caso se deseje que a reordenação seja varie entre experimentos.

Para realizar os testes no WEKA Foram utilizados os parâmetros (*cross-validation, percentage split e random seed*), foram realizados 30 testes, onde quinze foram feitos com *cross-validation* e quinze com *percentage split*. Os vinte testes do *cross-validation* foram subdivididos em quatro grupos de tamanhos iguais diferindo entre si no parâmetro *n*, sendo assim foram utilizados os seguintes valores: 10, 15, 20, 25 e 30. De modo análogo o mesmo foi feito com o *percentage split*, tendo como valores: 66%, 70%, 75%, 80% e 85%. Quanto ao parâmetro *random seed*, foram utilizados valores 1, 2 e 3 em cada sequencia de testes (*cross-validation e percentage split*). Os dados das classificações foram analisados levando em conta os parâmetros de Classificação Correta; Coeficiente de Kappa e número de erros e acertos mostrados pela Matriz de Confusão.

O Kappa é um coeficiente bastante utilizado para estudos de confiabilidade, porém apresenta limitações no que diz respeito a fornece informações da estrutura de concordância e discordância não considerando alguns aspectos importantes presentes nos dados (SILVA; PEREIRA, 1998). Desta forma é interessante incorporar outras abordagens visando complementar a análise (LANDIS; KOCH, 1977; KOTZ; JOHNSON, 1983; SILVA; PEREIRA, 1998).

A classificação correta dos dados é obtida através da porcentagem de acerto e significa quanto o algoritmo foi capaz de classificar de maneira correta as instâncias com relação ao número total. O Coeficiente Kappa apresenta-se como uma medida de concordância do quanto à classificação está de acordo com os dados a serem utilizados para descrever e testar o grau de concordância, confiabilidade e precisão na classificação (LANDIS; KOCH, 1977; KOTZ; JOHNSON, 1983; SILVA; PEREIRA, 1998).

Tabela 1. Atributos da Base de Dados

ATRIBUTOS DA BASE DE DADOS ORIGINAL WISCONSIN
1. Número do código da amostra: número id
2. Espessura do agrupamento: 1 – 10

3. Uniformidade do tamanho da célula: 1 – 10
4. Uniformidade da Forma da Célula: 1 – 10
5. Adesão Marginal: 1 – 10
6. Tamanho Único de Célula Epitelial : 1 – 10
7. Núcleos Nus: 1 – 10
8. Cromatina Branda : 1 – 10
9. Nucleoles Normais: 1 – 10
10. Mitoses: 1 – 10
11. Classe: (2 para benigno, 4 para maligno)

Fonte: Base de dados original do câncer de mama de Wisconsin, 1992.

Tabela 2. Interpretação da estatística kappa

Valor Kappa	Grau de Concordância
0	Nenhum
0-0,2	Leve
0,2-0,4	Bom
0,4-0,6	Moderado
0,6-0,8	Considerável
0,8-1,0	Quase Perfeito

Fonte: MCGINN et all, 2004.

RESULTADOS E DISCUSSÃO

Os resultados obtidos através dos testes foram organizados nas tabelas 3 e 4 de acordo com o pior e melhor resultado de cada método de separação de conjuntos de treinamento e de teste (*cross-validation, percentage Split e random seed*). Para cada grupo foram informados a porcentagem de acerto máximo, assim como o kappa encontrado seguidos de suas classificações, os detalhes dos testes encontram-se no CD.

Tabela 3. Resultados dos testes com *cross-validation.*

Folds	**Intervalo de acertos**	**Coeficiente Kappa**	**Classificação do Kappa**
10	95,99% - 96,13%	91,57%	Quase Perfeito
15	95,99% - 96,13%	91,57%	Quase Perfeito
20	95,99% - 96,13%	91,57%	Quase Perfeito
25	96,13% - 96,28%	91,89%	Quase Perfeito
30	96,13% - 96,13%	91,57%	Quase Perfeito

Tabela 4. Resultados dos testes com *percentage Split.*

Split	Intervalo de acertos	Coeficiente Kappa	Classificação do Kappa
66	95,37% - 97,05%	93,62%	Quase Perfeito
70	95,23% - 97,14%	93,92%	Quase Perfeito
75	94,85% - 97,14%	93,98%	Quase Perfeito
80	95,71% - 97,14%	95,06%	Quase Perfeito
85	95,23% - 97,85%	95,77%	Quase Perfeito

Diante da realização dos testes o que resultou no melhor kappa foi o *percentage split* com slip 85% e *random seed* 3 sendo coeficiente Kappa=95,77%, já o menor Kappa foi obtido através do teste *percentage split* com Split 75% e *random seed* 1, sendo coeficiente Kappa=89,09% em todos os testes realizados foi obtido um coeficiente Kappa quase perfeito.

Observou-se que os maiores coeficientes Kappa foram encontrados nos testes do *percentage Split*, porém em todos os testes realizados em todos os parâmetros o Kappa obteve classificação quase perfeito. Um estudo que comparou o Naive Bayes a outros algoritmos afirma que este algoritmo apesar de simplificado, diversas vezes funciona melhor em situações muito complexas do mundo real e que uma de suas principais vantagens é que o mesmo requer uma pequena quantidade de dados de treinamento para estimar os parâmetros (ALAM; PACHAURI, 2017).

CONSIDERAÇÕES FINAIS

O algoritmo Naive Bayes apresenta-se como um ótimo classificador de dados. Embora, em primeira instância, a suposição de que os atributos sejam independentes entre si seja considerados “inocente”, os testes realizados trouxeram resultados positivos além de ter apresentado agilidade na execução.

Dessa forma, tendo em vista que a base de dados analisada é da área de saúde, sugere-se que essa técnica seja empregada na área da ciência da saúde a fim de utilizar como classificador para auxiliar os gestores no processo de tomada de decisão para melhoria do sistema de saúde.

A utilização de recursos computacionais é altamente requerida, uma vez que a realização de cálculos só é possível com este recurso. Redes Bayesianas oferecem uma maneira conveniente de

atacar uma infinidade de problemas em que se deseja chegar a conclusões, não só de forma lógica, mas também probabilística.

REFERÊNCIAS

ALAM, F.; PACHAURI, S. Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA. **Advances in Computational Sciences and Technology** ISSN 0973-6107 Volume 10, Number 6 (2017) pp. 1731-1743. Disponível em:< http://www.ripublication.com/acst17/acstv10n6_19.pdf> Acesso em: 13 set 2018.

BOUCKAERT, R. R. et al. WEKA manual for version 3-8-1. **University of Waikato, New Zealand**, 2016.

CARVALHO, J. V. F.; CHIANN, C. Redes Bayesianas: Um Método para Avaliação de Interdependência e Contágio em Séries Temporais Multivariadas. **RBE**, Rio de Janeiro, v. 67, n. 2, p. 201–217, 2013.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. "Inteligência Artificial: Uma abordagem de aprendizado de máquina", **LTC**, p. 70, 2011.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações, 2. ed. **Elsevier**, 2015.

HOJSGAARD, S. et al. Graphical Models with R. New York: Springer, 2012.

KAYMAK, U.; BERG, J. van der. Conditional density estimation using probabilistic fuzzy systems. IEEE Transactions on Fuzzy Systems, 2013.

LANDIS, J. R.; KOCH, G. G. **The measurement of observer agreement for categorical data. Biometrics.**33:159-75,1977.

MARTINS, F. G.; COELHO, L. S. Aplicação do método de análise hierárquica do processo para o planejamento de ordens de manutenção em dutovias. **Revista GEPROS**, n. 1, p. 65, 2014.

MORAES, R. M.; MACHADO, L. S. ***Gaussian Naive Bayes for Online Training Assessment in Virtual Reality-Based Simulators***. Mathware & Soft Computing, v.16, n.2, p.123-132, 2009a.

PUURULA, A**.** Combining modi_cations to multinomial naive bayes for text classication**.** In Hou, Y., Nie, J.Y., Sun, L., Wang, B., Zhang, P., eds.: **Information Retrieval Technology**. Volume 7675 of Lecture Notes in Computer Science. Springer Berlin Heidelberg 114-125. 2012.

RAIZADA, R. D. S.; LEE, Y. S. Suavidade sem suavização: Por Gaussian Naive Bayes Não é ingênuo de Estudos Holofote Multi-Assunto. 2013. **PLoS ONE 8** : e69566.doi: 10.1371 / journal.pone.0069566.

RASCHKA, S. Naive bayes and text classification I-introduction and theory. **arXiv preprint arXiv**:1410.5329, 2014.

SÁ, J. M. C. et al. Análise de Crédito Utilizando uma Abordagem de Mineração de Dados. **Revista de Engenharia e Pesquisa Aplicada** (2018) Vol.3 No.3. Disponível em:< http://www.revistas.poli.br/index.php/repa/article/view/967/453> Acesso em: 10 set 2018.

SAPNA, JAIN.; M, AFSHAR AALAM AND M N DOJA, "K-means clustering using weka interface", Proceedings of the 4th National Conference; **INDIACom**-2010.

SILVA, M. P. S. Mineração de Dados – Conceitos, Aplicações e Experimentos com Weka. Mossoró-RN, Brasil: **Universidade do Estado do Rio Grande do Norte**, 2004.

UNIVERSITY OF WAIKATO. WEKA 3 – **Machine Learning Software in Java**. Disponível no site da University of Waikato. Disponível em: < http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 10 set 2018.

VIER, J. et al. Empregando Redes Bayesianas para modelar automaticamente o conhecimento dos aprendizes em Lógica de Programação. **Revista Brasileira de Informática na Educação**, Volume 23, Número 2, 2015.