O TEOREMA DE PROBABILIDADE PELO ALGORITMO NAIVE BAYES PARA A CLASSIFICAÇÃO DE DADOS: UM ESTUDO EM FERRAMENTAS DE DATA MINING

Marcio Novaski[1], Merisandra C. de Mattos Garcia[1,2], Maicon Bastos Palhano[1], Gabriel Felippe[1], Ruano Marques Pereira[1] e Fernando Mendes de Azevedo[2]

[1]*Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma –SC- Brasil*

[2]*Instituto de Engenharia Biomédica, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina Florianópolis-SC- Brasil*

RESUMO

Os constantes avanços tecnológicos têm permitido que as mais diversas organizações gerem repositórios de dados cada vez maiores tornando necessário o desenvolvimento de tecnologias que auxiliem nas análises e obtenção de conhecimento a partir destes dados. O *data mining* destaca-se dentre essas tecnologias utilizando algoritmos específicos para extração de possíveis padrões existentes nesses conjuntos de dados. Este artigo demonstra a implementação do algoritmo Naive Bayes, que utiliza o Teorema de Bayes e os conceitos de estatística e probabilidade para a tarefa de classificação, e a sua aplicação em uma base de dados da área médica referente á dados clínicos de pacientes com diagnóstico positivo para o câncer de mama. A tarefa de classificação consiste na identificação de propriedades comuns entre os elementos de uma base de dados e a associação desses elementos a uma classe predefinida, sendo que o algoritmo implementado atribui ao registro o rótulo de classe que apresentar a probabilidade máxima *a posteriori*, considerando a independência condicional entre os atributos, o que simplifica os cálculos necessários tornando mais rápido o processo de treinamento do algoritmo. Ao final da pesquisa foram realizados testes em uma base de dados e o resultados gerados pelo algoritmo foram avaliados usando algumas medidas de qualidade como sensibilidade, especificidade, acurácia, confiabilidade positiva e índice Kappa.

PALAVRAS-CHAVES

Inteligência Computacional, Data Mining, Classificação, Algoritmo Naive Bayes, Probabilidade.

1. INTRODUÇÃO

O *data mining* objetiva a descoberta de padrões válidos e novos em grandes conjuntos de dados (Fayyad et al, 1996), sendo composta por tarefas, no qual são implementadas em ferramentas conhecidas como shells, sendo que existe em desenvolvimento a Shell Orion Data Mining Engine, que consiste em um projeto implementado em âmbito acadêmico. Dentre as diversas tarefas existentes no processo de *data mining*, a Shell Orion atualmente disponibiliza as tarefas de associação, classificação e clusterização.

A classificação é vista como uma das mais populares e utilizadas tarefas do *data mining* (Goldschmidt e Passos, 2005). Consiste em associar um registro de uma determinada base de dados a uma classe predefinida. Os classificadores estatísticos baseados no teorema de Bayes, também conhecidos como classificadores bayesianos, utilizam a probabilidade como principal recurso para identificação da classe. Usados para lidar com situações de incerteza por aleatoriedade, destacam-se por apresentar um menor tempo no processo de aprendizagem.

Neste artigo apresenta-se o desenvolvimento do algoritmo Naive Bayes, baseado no teorema de Bayes, no módulo de classificação da Shell Orion e uma avaliação dos seus resultados por meio de medidas de validação em *data mining*, bem como a comparação dos resultados com o de outra ferramenta, no caso a Weka.

2. O ALGORITMO NAIVE BAYES NA TAREFA DE CLASSIFICAÇÃO DA SHELL ORION DATA MINING ENGINE

O algoritmo Naive Bayes foi proposto por Richard Duda e Peter Hart no livro Pattern Classification and Scene Analisys em 1973 na Califórnia, Estados Unidos, originado do trabalho de reconhecimento de padrões e análise de cenas em aprendizado de máquina.

É um classificador estatístico que utiliza os conceitos da probabilidade para identificar a qual classe predefinida pertence um determinado registro. A sua principal característica é a suposição da independência condicional, ou seja, ele assume que o efeito do valor de um atributo a uma determinada classe é independente dos valores dos demais atributos (Coppin, 2010).

A modelagem do algoritmo Naive Bayes iniciou-se com a construção dos diagramas de caso de uso, atividades e seqüência utilizando os padrões Unified Modeling Language (UML). Posteriormente foi desenvolvida a demonstração matemática do funcionamento do algoritmo com a finalidade de facilitar o entendimento e a sua implementação.

O algoritmo necessita de dois conjuntos de dados sendo um para o treinamento e outro para testes. Definidos os dois conjuntos de dados, o primeiro passo consiste em:

a) **identificar o número de classes;**

b) **identificar o total de amostras no conjunto de treinamento;**

c) **identificar o total de amostras para cada classe;**

Com os parâmetros obtidos, calcula-se primeiramente a probabilidade *a priori* de cada classe conforme a fórmula:

$$P(C_i) = \frac{S_i}{S}, i = 1, \dots, n$$

Esse cálculo consiste na divisão do número de amostras S_i rotuladas como classe C_i pelo total de amostras S. Visto que o cálculo da probabilidade é dado por uma fração, se o numerador assumir o valor zero o resultado pode ser comprometido. Para evitar esse problema utiliza-se o teorema da aproximação de Laplace adicionando a constante 1 ao numerador e a variável k ao denominador:

$$\frac{n_c + 1}{n + k}$$

Onde n_c é o número de instâncias pertencentes à classe C_i, n é o número total de instâncias de treinamento com classe C_i e k corresponde à quantidade máxima de valores distintos, ou seja, k=3.

A fórmula para o cálculo da probabilidade *a priori* de cada classe é alterada para:

$$P(C_i) = \frac{S_i + 1}{S + k}, i = 1, \dots, n$$

O próximo passo consiste em calcular a probabilidade de cada atributo da amostra de teste condicionada a cada uma das classes aplicando o teorema de Laplace à fórmula:

$$P(x_k|C_i) = \frac{S_{ik} + 1}{S_i + k}$$

Estima-se então as probabilidades da amostra total $P(X|C_i)$. Segundo Coppin (2010), Duda e Hart (1973) e Han e Kamber (2001), para reduzir o custo computacional durante o cálculo de $P(X|C_i)$, o Naive Bayes considera a independência condicional de classe, ou seja, o efeito do valor de um determinado atributo sobre uma classe é considerado independente dos valores dos outros atributos. Esta afirmação é dada por:

$$P(X|C_i) = \prod_{k-1}^{n} P(x_k|C_i)$$

Posteriormente as regras de classificação são aplicadas considerando-se a probabilidade *a priori* de cada classe:

$$P(X|C_i) \times P(C_i)$$

O resultado obtido pela a aplicação das regras de classificação define a máxima *a posteriori* que indicará a qual classe pertence o registro analisado.

2.1 Implementação

O algoritmo Naive Bayes foi desenvolvido no módulo de classificação da Shell Orion Data Mining Engine, por meio da linguagem de programação Java e do ambiente de programação integrado NetBeans 7.1.1.

Para a execução do algoritmo é necessário que seja informado pelo usuário o método de teste a ser realizado. Entre as opções disponíveis na Shell Orion tem-se o método *holdout* e a opção de usar o conjunto de treinamento.

A opção *holdout* permite resultados mais confiáveis, uma vez que as amostras reservadas para o conjunto de testes não fazem parte do conjunto de treinamento (Witten et al, 2011).

A sugestão de proporção para dados reservados é de 2/3 para treinamento e 1/3 para testes, mas a alteração do percentual de dados reservado para treinamento também é permitida (Tan et al, 2009).

Nas análises dos resultados obtidos, foram empregados os seguintes medidas de qualidade:

a) **sensibilidade:** é a capacidade do classificador para encontrar os registros que realmente pertencem a classe considerada;
b) **especificidade:** consiste na capacidade do classificador para encontrar os registros que não pertencem a classe considerada;
c) **acurácia:** é o grau de exatidão, relação entre os valores estimados e os valores reais;
d) **erro:** refere-se à taxa de erro geral, ou seja, a proporção de registros classificados incorretamente;
e) **confiabilidade positiva:** é a capacidade do classificador para identificar corretamente os verdadeiros positivos;
f) **índice kappa:** é o coeficiente de avaliação da concordância entre os resultados.

2.2 Resultados Obtidos

A base de dados utilizada na avaliação do algoritmo Naive Bayes na Shell Orion refere-se a dados clínicos de pacientes com diagnóstico positivo para o câncer de mama dos hospitais da Universidade de Wisconsin, Madison nos Estados Unidos disponibilizada pelo repositório UCI *Machine Learning Repository*. É composta por 683 registros com 10 atributos e 2 classes que identificam o resultado do diagnóstico como benigno ou maligno.

Os testes realizados abrangem a análise da precisão dos resultados obtidos e a comparação com os resultados da ferramenta Weka 3.6.7 (http://www.cs.waikato.ac.nz/ml/weka/), a comparação entre os tempos de processamento durante o treinamento e a classificação na Shell Orion, e os tempos de processamento entre a Shell Orion e a ferramenta Weka.

2.2.1 Classes Identificadas pelo algoritmo Naive Bayes

O teste para a identificação das classes e validação do desempenho foi realizado utilizando o método *holdout*. Optou-se por repetir o teste por três vezes usando percentuais diferentes para cada teste.

A melhor taxa de acertos confirmou que a proporção de 2/3 dos registros reservados para treinamento apresenta resultados mais precisos.

A tabela 1 mostra a matriz de confusão que exibe o número de classificações preditas versus o número de classificações corretas. Por meio da contagem destes registros tem-se a avaliação de desempenho do modelo (Tan et al, 2009).

Tabela 1. Matriz de confusão – Classificação da base do câncer de mama na Shell Orion

Classe	Predita Classe 2	Predita Classe 4	Total
Verdadeira Classe 2	176	2	178
Verdadeira Classe 4	0	54	54
Total	176	56	232

Dos valores obtidos na matriz de confusão pode-se extrair as seguintes variáveis mostradas na tabela 2:

Tabela 2. Relação de verdadeiros e falsos positivos e verdadeiros e falsos negativos.

Variável	Valor
Verdadeiros Positivos	176
Falsos Positivos	0
Verdadeiros Negativos	54
Falsos Negativos	2

Com os valores das variáveis da tabela 2, calcularam-se as medidas de qualidade do modelo. A tabela 3 demonstra os valores obtidos pela Shell Orion.

Tabela 3. Medidas de Qualidade – Shell Orion.

Variável	Shell Orion
Registros classificados corretamente	230
Registros classificados incorretamente	2
Sensibilidade	98,88%
Especificidade	100%
Acurácia	99,14%
Erro	0,86%
Confiabilidade positiva	100%
Índice Kappa	0,976

A análise dos valores obtidos nas medidas de qualidade mostram que o desempenho do módulo desenvolvido pode ser considerado satisfatório por apresentar medidas próximos de 100% e taxa de erro próxima de zero.

2.2.2 Comparação das Medidas de Qualidade com a Ferramenta Weka 3.6.7

Nessa etapa analisou-se a qualidade dos resultados gerados pela Shell Orion e comparou-se com os resultados obtidos pela ferramenta Weka. Apenas as medidas de qualidade foram considerados nessa avaliação, portando os tempos de processamento foram ignorados.

A análise seguinte optou-se por utilizar o método *holdout* reservando 66% da base de dados para treinamento e o restante para teste. Os valores obtidos para as medidas de qualidade nas duas ferramentas são mostrados na tabela 7.

Tabela 6. Medidas de qualidade – Comparação entre Shell Orion e Weka

Variável	Shell Orion	Weka
Registros classificados corretamente	230	229
Registros classificados incorretamente	2	3
Sensibilidade	98,88%	98,88%
Especificidade	100%	98,1%
Acurácia	99,14%	98,7%
Erro	0,86%	1,3%
Confiabilidade positive	100%	99,4%
Índice Kappa	0,976	0,964

A análise dos valores obtidos nas medidas de qualidade e a comparação com a Weka mostrou que a Shell Orion obteve resultados melhores em quase todos as medidas confirmando o seu correto funcionamento, especialmente no que diz respeito a percentuais de acerto (Shell Orion – 99,14%; Weka – 98,71%) e índice Kappa (Shell Orion – 0,976; Weka – 0,964).

Concluiu-se que ambas as ferramentas têm comportamento semelhante com relação à qualidade dos resultados do processo de classificação.

3. CONCLUSÃO

Este artigo apresentou o algoritmo Naive Bayes que utiliza os conceitos de estatística e probabilidade baseados no teorema de Bayes, para a tarefa de classificação da Shell Orion Data Mining Engine, contribuindo com a ampliação das funcionalidades da ferramenta.

Pode-se confirmar, diante dos resultados obtidos, a aplicabilidade do algoritmo para a tarefa de classificação visto que os resultados avaliados apresentaram boas medidas de qualidades relacionados à precisão e confiabilidade dos valores obtidos.

Concluiu-se que o algoritmo foi implementado com sucesso, pois apresentou resultados satisfatórios obtidos no processo de classificação e ainda melhores quando comparados com uma ferramenta de *data mining* bastante utilizada e também gratuita, confirmando o seu correto funcionamento.

REFERÊNCIAS

Coppin, B. 2010. *Inteligência artificial*. LTC, Rio de Janeiro.

Fayyad, U. M., Piatetsky-shapiro, G. e Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI Magazine*, Providence, v.17, n. 3, p. 37-54.

Duda, R. O., Hart, P. E. e Stork, D. G. 2000. *Pattern Classification*, JohnWiley & Sons, New York.

Goldschmidt, R. e Passos, E. L. 2005. *Data mining: uma guia prático*, Elsevier, Rio de Janeiro.

Han, J. e Kamber, M. 2006. *Data mining: concepts and techniques.* Morgan Kaufmann, San Francisco.

Tan, P. et al. 2009. *Introdução ao Datamining: mineração de dados*, Ciência Moderna, Rio de Janeiro.

Witten, I. H. et al. 2011. *Data mining practical machine learning tools and techniques,* Morgan Kaufmann, Burlington.