A INTELIGÊNCIA COMPUTACIONAL E O TEOREMA DE THOMAS BAYES, UMA DAS FERRAMENTAS DO SOFTWARE R

MARLON FERREIRA CORSI

Mestrado em Modelagem Matemática e Computacional pela Universidade

Federal do Rio de Janeiro (UFRRJ), Rio de Janeiro, RJ, Brasil.

ofmarlon@yahoo.com.br

BRUNO AREAL DE SANTANA

Mestrado em Modelagem Matemática e Computacional pela Universidade

Federal do Rio de Janeiro (UFRRJ), Rio de Janeiro, RJ, Brasil.

profbfisica@hotmail.com

RESUMO

O artigo mostra a implementação de uma rede bayesiana utilizando o software “*The R Project for Statistical Computing*” a fim de estimar as probabilidades a partir de dados hipotéticos, de um problema real, que trata os aspectos de um grupo de estudantes que obtiveram aprovação ou reprovação visto que estes ou fizeram curso ou possuíam internet banda larga em suas residências. Estes dados foram tratados no software NETICA® e os resultados obtidos pelo modelo criado no “R” será comparado com os desse software. Mostrando a aplicabilidade de um programa open source como alternativa do programa comercial HUGIN® com licença estudante. Este será usado apenas para fazer a inferência da Rede bayesiana.

Palavras-chave: Teorema de Bayes. Software R. Hugin.

INTELLIGENCE COMPUTER AND THOMAS BAYES THEOREM, ONE OF THE TOOLS SOFTWARE R

ABSTRACT

The article shows the implementation of a Bayesian network using “The R Project for Statistical Computing” to estimate the odds from hypothetical data, a real problem, which deals with aspects of a group of students who have obtained approval or disapproval as these or made progress or had broadband internet in their homes. These data were processed in NETICA® software and the results of the model created in the “R” will be compared with that software. Showing the applicability of an open source alternative to the commercial program HUGIN® licensed student. This will only be used to make the inference of Bayesian Network.

Keywords: Bayes Theorem. Software R. Hugin.

1 INTRODUÇÃO

A Rede Bayesiana tem origem no teorema de Thomas Bayes, constituindo um modelo gráfico presente na computação, que utiliza os fundamentos matemáticos da probabilidade e da estatística acoplados a teoria dos grafos. A teoria define as dependências causais entre as variáveis de maneira clara, de forma gráfica formando um conjunto de relações de causa e efeito. Em uma aplicação sobre um domino real, uma rede bayesiana interpreta os fatos na qual é submetida e estima uma resposta levando em conta o aprendizado da rede. A metodologia emprega uma abordagem interpretativa e analítica para a inteligência artificial probabilística, e aplicada nas áreas de estimação de risco operacional, diagnostico medico, analise de credito, projeto de jogos computacionais, imputação de dados, entre outras, o que a torna uma ferramenta essencial pra fins de decisão, classificação de variáveis.

O cálculo das probabilidades são baseados em fenômenos de observação, neces-

sitam de modelos matemáticos (determinísticos ou probabilísticos) dos quais simplificam seu entendimento. A fim de verificar a validade de um modelo, deveremos deduzir um certo número de conseqüências do modelo e, a seguir, comparar esses resultados previstos com as observações. É muito importante distinguir a própria observação do modelo que o representa, por isso é necessária a utilização de modelos apropriados a circunstância dos experimentos. Para isso, temos modelos determinísticos e não determinísticos (probabilísticos). A correlação entre a teoria das probabilidades e a teoria dos grafos compõem os modelos gráficos.

A teoria dos grafos estudam objetos combinatórios e modelam muitos problemas da matemática, das engenharias, indústria e informática. Um grafo G consiste de dois conjuntos V e E . O conjunto V é um conjunto de nós finito e não vazio. O conjunto E é um conjunto de pares de nós, estes pares são chamados de arestas (arcos). A notação $V(G)$ e $E(G)$ representa os conjuntos de nós e arestas, respectivamente, do grafo G . Escreve-se também $G=(V,E)$ para representar um grafo.

Figura 1 - Exemplos de Grafos

2 METODOLOGIA

A partir do espaço amostral, foram determinados os conjuntos de fatos de interesses, e estimada as probabilidades condicionais entre os eventos. O espaço amostral é apresentado na Tabela 1.

Tabela 1 - Dados de trabalho

Candidato (Nº Sequencial)	Aprovado	Cursinho	IBL
1	Sim	Nao	Sim
2	Nao	Sim	Nao
3	Sim	Sim	Nao
4	Nao	Nao	Nao
5	Sim	Sim	Nao
6	Sim	Sim	Sim
7	Nao	Sim	Nao
8	Nao	Sim	Nao
9	Sim	Nao	Sim
10	Nao	Sim	Nao
11	Sim	Sim	Nao
12	Nao	Nao	Nao
13	Sim	Sim	Nao
14	Sim	Sim	Sim
15	Nao	Sim	Nao
16	Nao	Sim	Nao
17	Sim	Nao	Sim
18	Nao	Sim	Nao
19	Sim	Sim	Nao
20	Nao	Nao	Nao
21	Sim	Sim	Nao
22	Sim	Sim	Sim
23	Nao	Sim	Nao
24	Nao	Sim	Nao

A partirdo espaço amostral foram construídas duas tabelas que representam quantitativamente a pesquisa realizada, as tabelas 2 e 3.

Tabela 2 - Aprovados e Reprovados

	IBL	IBL/CURSO	CURSO
REP	0	0	9
AP	3	3	6

Tabela 3 - Metodologia de Estudo

	IBL	IBL/CURSO	CURSO
SIM	3	3	15
NÃO	16	2	4

A partir dos dados foram estimadas as probabilidades condicionais, que consiste em tratar a probabilidade de um evento **A** ocorrer dado que um evento **B** também ocorrer, ambos pertencentes ao mesmo espaço amostral, porém a probabilidade de **A** é calculada sobre o evento **B** e não sobre o espaço amostral. Para calculá-la utiliza-se a fórmula:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A)P(B|A)$$

Um evento **A** é considerado independente de um outro evento **B** se a probabilidade de **A** é igual à probabilidade condicional de **A** dado **B**, isto é, se $P(A) = P(A|B)$. Evidente que, se **A** é independente de **B**, **B** é independente de **A**, assim: $P(B) = P(B|A)$. Considerando o teorema do produto, pode-se afirmar que se **A** e **B** são independentes, então: $P(A \leftrightarrow B) = P(A) \cap P(B)$.

A probabilidade de um evento **A** ocorre dados dois outros eventos **B** e **C** é:

Foram estimadas as probabilidades condicionais considerando que um aluno foi aprovado dado que este tenha feito curso e possua internet banda larga na sua residência.

Tabela 4 - Percentual

		P(A C,I)	
CURSO	IBL	SIM	NÃO
SIM	SIM	1	0
SIM	NÃO	0,4	0,6
NÃO	SIM	1	0
NÃO	NÃO	0	1

A partir da tabela de percentual, foi criada uma rede bayesiana no programa "R" de forma a se obter o valor que justifique o resultado de APROVADO relacionado com as variáveis de entrada de CURSO e IBL.

O script implementado em "R":

```
library(RHugin)
```

```
rede<- hugin.domain()
```

```
add.node(rede,"Aprovado",states = c("Sim","Nao"))
```

```
add.node(rede,"Curso",states = c("Sim","Nao"))
```

```
add.node(rede,"IBL",states = c("Sim","Nao"))
```

```
add.edge(rede,"Aprovado","Curso")
```

```
add.edge(rede,"Aprovado","IBL")
```

```
plot(rede)
```

```
curso.table<- get.table(rede,"Curso")
```

```
curso.table[["Freq"]] <-c(0.75, 0.25)
```

```
set.table(rede,"Curso",curso.table)
```

```
ibl.table<- get.table(rede,"IBL")
```

```
ibl.table[["Freq"]] <-c(0.25, 0.75)
```

```
set.table(rede,"IBL",ibl.table)
```

```
aprovado.table<- get.table(rede,"Aprovado")
```

```
aprovado.table[["Freq"]] <-c(1.0, 0.0, 1.0, 0.0, 0.4, 0.6, 0.0, 1.0 )
```

```
set.table(rede,"Aprovado",aprovado.table)
```

```
get.table(rede,"Aprovado", class ="ftable")
```

```
write.rhd(rede, file= "vestibular.net",type = "net")
```

3 RESULTADOS

A rede foi criada no “R” porém a inferência foi realizada no Hugin, um programa comercial que disponibiliza licença estudante, porém com limite de nó da rede. Ao gerar a rede, está é salva e aberta no Hugin, mesmo que tenha uma quantidade maior de nós do que a licença estudante permite. Foi utilizada uma rede simples, que a partir do código é graficamente representada pela Figura 2.

Figura 2 – Rede Bayesiana

O programa Hugin oferece como resposta o gráfico da Figura 3.

Figura 3 – Curva ROC

O relatório de análise oferecido pelo Hugin está presente na Figura 4, onde o programa utiliza a tabela de entrada para estimar os resultados, o valor selecionado para a quantidade de casos foi 100.

Figura 4 – Relatório de Análise

Analysis Report

Node: Aprovado
Number of cases: 100

Using cutoff threshold = Max. belief

Confusion Matrix:

[actual]

	Sim	Nao [predicted]	
31	0	Sim	
11	58	Nao	

Error rate: 11.00

Avg. Euclidian distance: 0.17611

Avg. Kulbach-Leibler divergence: 0.27540

No caso apresentado a taxa de erro foi de 11% e a curva ROC se apresentou graficamente satisfatória, estando bem distante da diagonal.

4 CONSIDERAÇÕES FINAIS

O artigo apresentou uma alternativa para criação de redes bayesianas, utilizando o software “*The R Project for Statistical Computing*”, o qual mostrou ser uma ferramenta muito vantajosa, tendo em vista a sua aplicabilidade. A integração de um *software open source* com um *software* comercial Hugin@ versão shareware viabilizou bastante, visto que não houve custo com essas aquisições e seus resultados foram satisfatórios para um modelo preditivo visto que a taxa de erro da rede foi muito baixa.

REFERÊNCIAS

BORDEN, Richard L. e FAIRES, J. Douglas. **Numerical Analysis**. Brooks/cole: 2010, 9th Edition.

CASTRO, L.N. **Fundamentals of Natural Computing: Basic Concepts, Algorithms and Applications** CHAPMAN & HALL, 2006.

GOLDSCHMIDT, R.R., **Uma Introdução a Inteligência Computacional – Fundamentos, Ferramentas e Aplicações** – ISTRJ – Série Livros Didáticos Digitais, 2010.

HUDSON,D, COHEM,M. **Neural Network and Artificial Intelligence for Biomedical Engi-**

neering, 1996.

Software R “The R Project for Statistical Computing” (<http://www.r-project.org/>).

SZWARCFITER, Jayme Luiz e MARKENZON Lilian. **Estruturas de Dados e Seus Algoritmos**. LTC:2010, 3ª Edição.