UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS CORNÉLIO PROCÓPIO
DIRETORIA DE GRADUAÇÃO E EDUCAÇÃO PROFISSIONAL
DEPARTAMENTO DE COMPUTAÇÃO
ENGENHARIA DE COMPUTAÇÃO

RAFAEL WILSON DANTAS DA SILVA

CONSTRUÇÃO DE ALGORITMOS BAYESIANOS UTILIZANDO ALGORITMOS EVOLUTIVOS

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO
2017

RAFAEL WILSON DANTAS DA SILVA

CONSTRUÇÃO DE ALGORITMOS BAYESIANOS UTILIZANDO ALGORITMOS EVOLUTIVOS

Trabalho de Conclusão de Curso apresentada ao Departamento de Computação da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de “Bacharel em Engenharia de Computação”.

Orientador: Prof. Dr. Danilo Sipoli Sanches

Co-orientador: Prof. Dr. Carlos Nascimento Silla Jr.

CORNÉLIO PROCÓPIO
2017

RESUMO

SILVA, Rafael. **Construção de algoritmos bayesianos utilizando algoritmos evolutivos**. 2017. 17 f. Trabalho de Conclusão de Curso – Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2017.

Texto do resumo (máximo de 500 palavras).

Palavras-chave: Palavra-chave 1. Palavra-chave 2. (entre 3 e 5 palavras)

ABSTRACT

SOBRENOME, Nome. **Construction of Bayesian algorithms using evolutionary algorithms.** 2017. 17 f. Master Thesis – Electrical Engineering Graduate Program, Federal University of Technology - Paraná. Cornélio Procópio, 2017.

This is the english abstract. (maximum of 500 words).

Keywords: Keyword 1. Keyword 2. (entre 3 e 5 palavras)

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACO	Ant Colony Optimization
KDBC	*K-Dependence Bayesians Classifier*
NB	*Naive Bayes*

SUMÁRIO

1 INTRODUÇÃO

Nos tempos atuais estamos sobrecarregados de dados. A quantidade de dados no mundo e em nossas vidas tende a crescer cada vez mais e não há nenhum sinal de que esta estimativa reduza. Os computadores atuais tornam o processo de armazenamento de dados muito simples para descartarmos qualquer que seja o dado produzido no cotidiano de nossos dias. Há uma grande diferença entre a quantidade de dados que produzimos e a quantidade de informação que conseguimos retirar dela (WITTEN; FRANK; HALL, 2017).

Mineração de dados é sobre resolver problemas por meio da análise de dados já presentes na base de dados e um dos seus objetivos é a descoberta de conhecimento através de técnicas computacionais, que são capazes de explorar um grande conjunto de dados evidenciando padrões e auxiliando na descoberta de conhecimento. Este desenfreado crescimento de base de dados, traz a mineração de dados para o primeiro plano das novas tecnologias (WITTEN; FRANK; HALL, 2017).

A contrução de um classificador é uma tarefa básica na analise de dados e reconhecimento de padrões na Mineração de Dados. Simplificadamente, um classificador é uma função que assemelha com base em um conjunto de atributos um valor ao rotulo da classe (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

As Redes Bayesianas (serão explicadas na seção 2.2) tem se destacado como uma das abordagens mais promissoras no processo de descoberta de conhecimento em base de dados e uma das suas implementações mais conhecidas é o classificador *Naive Bayes* (NB)(SEBASTIANI; ABAD; RAMONI, 2010). O Naive Bayes é um classificador probabilistisco e tem como principal pressuposto a independência entre seus atributos.

1.1 PROBLEMA

No situações reais a independência de atributos é irreal. Diante da limitação de independência entre os atributos, diversos trabalhos foram realizados sobre possíveis modificações na estrutura do NB. Um dos metódos de melhoria dos resultados obtidos pelo NB é extenção da estrutura do classificador, que consiste na represntação de dependências entre os atributos (JIANG et al., 2007).

Uma forma de extensão dos classificadores Bayesianos foi apresentada por Sahami (1996), a fim de lidar com as suposições de independência entre os atributos do NB, onde ele introduz o termo *K-Dependence Bayesians Classifier* (KDBC), onde o valor K representa o número máximo de nós pais que um atributo pode ter além da classe. Dessa forma um 0-KDBC seria o equivalente a um Naive Bayes (FLORES et al., 2011).

Em (COOPER, 1990) foi provado que tentar encontrar a melhor rede bayesiana gerada pelo KDBC é um problema que se encontra no domínio dos NP-Completos, ou seja, é inviável computacionalmente tentar encontrar a melhor solução.

1.2 JUSTIFICATIVA

2 CLASSSIFICAÇÃO

2.1 TEOREMA DE BAYES

Se A, B e C são evento, a probabilidade de que estes eventos ocorram, $P(A)$, $P(B)$ e $P(C)$ respectivamente, pode ser representada por um número real entre 0 e 1.

A probabilidade de que um evento está ligada diretamente a quantidade de vezes em que ele ocorre em relação a quantidade de vezes em que um experimento é realizado. Logo, se um evento E ocorre N vezes em um experimento realizado M vezes, então temos que $P(E) = M/N$.

Para entender melhor o Teorema de Bayes que será explicado neste capitulo considere as seguintes definições:

- Eventos mutualmente exclusivos

  Eventos mutualemente exclusivos são eventos que não podem ocorrer ao mesmo tempo. Se E e F são mutualmente exclusivos, então a probabilidade da união entre eles é dada pela seguinte equação:

  $$P(E \cup F) = P(E) + P(F) \tag{1}$$

  Da mesma forma podemos ter eventos que não sejam mutualmente exclusivos, ou seja, podem ocorrer ao mesmo tempo. Neste caso eles são representados pela seguinte equação:

  $$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{2}$$

  Onde, $P(E \cap F)$ é a verificação de que os eventos sejam independentes, definida pela equação 3:

  $$P(E \cap F) = P(E) \cdot P(F) \tag{3}$$

- Probabilidade Condicional

  A probabilidade condicional é a probabilidade de um evento E ocorrer em função da ocorrência de outro evento F. A sua representação é dado por $P(E|F)$ e é lida como “a probabilidade de E, dado que F é verdadeiro”. Uma das formas de calcular a probabilidade condicional entre os eventos E e F é dada pela seguinte equação:

  $$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{4}$$

  E da mesma forma para calcularmos $P(F|E)$, temos:

  $$P(F|E) = \frac{P(E \cap F)}{P(E)} \tag{5}$$

- Regra do Produto de Probabilidades

Se aplicarmos a multiplicação cruzada nas equações 4 e 5, obtemos a seguinte equação:

$$P(E \cap F) = P(E|F) \cdot P(F) = P(F|E) \cdot P(E) \tag{6}$$

Manipulando a Equação 6, igualando os dois termos equivalentes à direita, e rearranjando obtêm-se o Dadas as considerações das a cima, manipulando os dois termos mais a direita da equação 6 chegamos enfimm na equação do **Teorema de Bayes** como pode ser visto a seguir:

$$P(E|F) = \frac{P(F|E) \cdot P(E)}{P(F)} \tag{7}$$

Com base no Teorema de Bayes surgiram os classificadores Bayesianos que serão explicados nas seções 2.2 e 2.3.

2.2 REDES BAYESIANAS

Segundo (SEBASTIANI; ABAD; RAMONI, 2010), atualmente as redes Bayesianos são uma das abordagens mais promissoras para o processo de descoberta do conhecimento.

As redes bayesianas pertencem a uma classe mais geral de modelos chamados de modelos probabilísticos gráficos que surgiram da combinação da teoria de grafos e da teoria da probabilidade. O seu sucesso se deve a sua capacidade de lidar com modelos probabilísticos complexos atravéz da decomposição em componentes menores e acessiveis. Um modelo probabilístico gráfico é definido por uma grafo onde os nós representam variáveis estocasticas e os arcos representam as dependências entre tais variáveis. Esses arcos são marcos pela probabilidade de distribuição de interação entre as variáveis vinculadas(SEBASTIANI; ABAD; RAMONI, 2010).

Uma rede Bayesiana é um modelo gráfico probabilístico onde o grafo conectando suas variáveis é um Grafo Acíclico Dirigido (do inglês, DAG - *Directed Acyclic Graph*). Este grafo gerado representa suposições de independência condicional entre os atributos que são usados para fatorar a distribuição de probabilidade conjunta das variáveis da rede. Assim o processo de aprendizagem em bancos de dados, com grandes quantidades de dados torna-se possível.(SEBASTIANI; ABAD; RAMONI, 2010) .

2.3 CLASSIFICADORES BAYESIANOS

Um dos classificadores bayesianos mais conhecidos é o Naive Bayes (NB). Este classificador aprende do conjunto de treinamento a probabilidade condicional de cada atributo A devida ao rótulo da classe C. A classificação é feito então aplicando o teorema de Bayes para calcular a probabilidade da classe C, dada a instância de A_1,..., A_n e então prediz a classe com a maior probabilidade a posteriori. Este cálculo é baseado em uma suposição de independência onde os atributos A_i são condicionalmente independentes dado o valor da classe C (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

Quando representado por uma rede Bayesiana a estrutura do NB pode ser visto na figura 1. Essa rede consegue representar o principal pressuposto por trás do classificador NB, de que os atributos são independentes entre si e dependem apenas da classe.

Naive Bayes é um dos mais efeivos classificadores, levando em conta o seu desempenho, atualmente. O desempenho do NB é surpreendente dado que a suposição de independência entre os atributos é quase sempre irreal (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

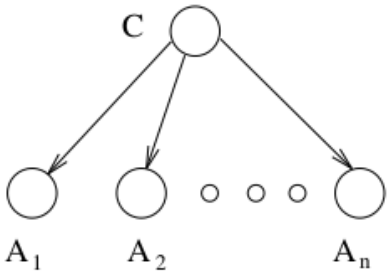


Figura 1 – Estrutura da rede do Naive Bayes

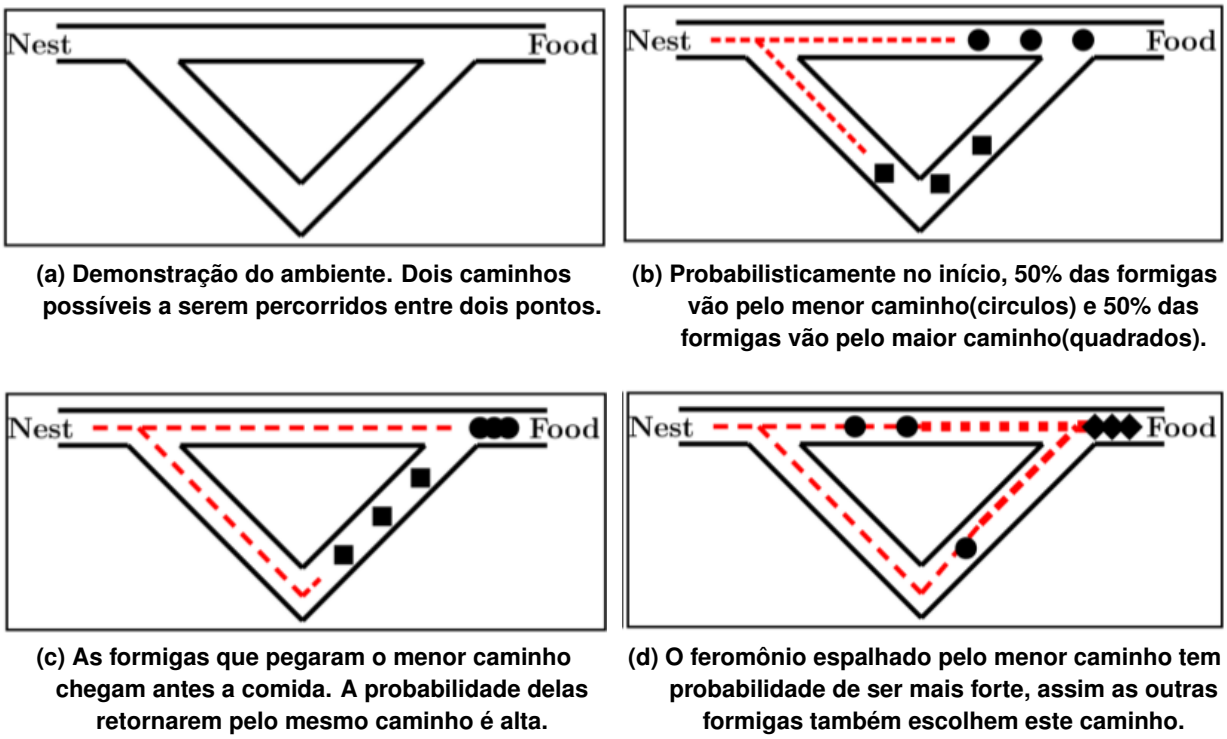
Fonte: (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

2.4 K-DEPENDENCE BAYESIAN CLASSIFIER (KDBC)

3 OTIMIZAÇÃO POR COLÔNIA DE FORMIGAS

Introduzido por Marico Dorigo e colegas no começos dos anos de 1990 (DORIGO; MANIEZZO; COLORNI, 1996)(DORIGO; BLUM, 2005), o Ant Colony Optimization (ACO)(em português, Otimização por Colônia de Formigas ou Otimização Colônia de Formigas) é baseado no comportamento das formigas ao se organizarem para encontrar o melhor percurso entre dois pontos, como por exemplo o lugar onde buscam comida e o ninho.

Inicialmente as formigas percorrem a área em busca de comida de forma randômica. Enquanto se movem as formigas deixam pelo caminho um rastro de ferômonio, que pode ser percebido pelas outras formigas. Quando a formiga encontra uma fonte de alimento, ela avalia a qualidade e quantidade do alimento e carrega o alimento de volta ao ninho. No caminho de retorno o feromônio produzido pela formiga pode depender da qualidade e da quantidade de comida encontrada e como a formiga que encontrou o melhor caminho retornará antes, o rastro deixado por ela será mais forte. Probabilisticamente as formigas tendem a seguir o caminhos com grandes quantidades de feromônio. Assim as outras formigas serão guiadas pelo melhor caminho (DRÉO; SIARRY, 2002) (BLUM, 2005). A demonstração do problema pode ser visto na figura **??** .



(a) Demonstração do ambiente. Dois caminhos possíveis a serem percorridos entre dois pontos.

(b) Probabilisticamente no início, 50% das formigas vão pelo menor caminho(circulos) e 50% das formigas vão pelo maior caminho(quadrados).

(c) As formigas que pegaram o menor caminho chegam antes a comida. A probabilidade delas retornarem pelo mesmo caminho é alta.

(d) O feromônio espalhado pelo menor caminho tem probabilidade de ser mais forte, assim as outras formigas também escolhem este caminho.

Figura 2 – Demonstração da capacidade da colonia de formigas em encontrar o melhor caminho entre dois pontos*Nest* e *Food*

Fonte: (BLUM, 2005)

4 ALGORITMOS EVOLUCIONÁRIOS

Algoritmos Evolucionários baseiam-se no processo de evolução natural proposto por Charles Darwvin (1850) para criar modelos computacionais na resolução de problemas. Processo este que mantem uma populão de individuos ou cromossomos que se comporta de forma semelhante à evolução das espécies.Cada indivíduo recebe uma avaliação que é quantificada em relação a solução do problema em questão, para então ser aplicado os operadores genéticos de forma a simular a sobrevivência dos indivíduos no meio (LINDEN, 2008).

Há uma grande variedade de modelos computacionais propostos na literatura, porém a grande maioria deles tem em comum a aplicação dos conceitos de seleção, mutação e reprodução na simulação da evolução das espécies. Estes processos dependem do desempenho dos indivíduos de cada espécie dentro do ambiente (LINDEN, 2008).

4.1 ALGORITMOS GENÉTICOS

Os algortimos Genéticos são técnicas heuristicas de otimização global. Algoritmos Genéticos são eficientes em buscar, no espaço das soluções, as que sejam tão próximas da solução ótimo quanto possível e isso quase sempre sem a interação humana. Portanto os algoritmos Genéticos são uma técnica adequada para problemas especialmente difíceis, como os problemas denominados NP completos (LINDEN, 2008).

5 PROPOSTA

5.1 METODOLOGIA DE PESQUISA

5.2 TRABALHOS RELACIONADOS

5.3 ATIVIDADES PLANEJADAS

Durante este trabalho serão realizadas as atividades descritas na tabela 1:

Tabela 1 – Atividades

Número	Descrição
1	Definição do tema para defesa da proposta
2	Entendimento do problema
3	Pesquisas relacionadas a Classificadores Bayesianos
4	Pesquisas relacionadas ao KBDC
5	Pesquisas relacionadas a Algoritmos Genéticos
5	Pesquisas relacionadas a Algoritmos de Busca como ACO
6	Verificação do estado da arte
7	Desenvolvimento da proposta
8	Implementação do Algoritmo Naive Bayes
9	Implementação do Algoritmo KDBC
10	Representação do KBDC em um Algortimo de Busca
11	Integração de Grafos e Método escolhido para o problema
12	Implementação Método escolhido para resolver o problema
13	Construção de sistema de integração entre os Algoritmos
14	Escrita da Monografia

5.4 CRONOMAGRAMA PLANEJADO

Tabela 2 – Cronograma de atividades

Atividades	Meses								
	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									

6 CONSIDERAÇÕES FINAIS

REFERÊNCIAS

BLUM, Christian. **Ant colony optimization: Introduction and recent trends**. 2005. Citado na página 12.

COOPER, Gregory F. The computational complexity of probabilistic inference using bayesian belief networks (research note). **Artif. Intell.**, Elsevier Science Publishers Ltd., Essex, UK, v. 42, n. 2-3, p. 393–405, mar. 1990. ISSN 0004-3702. Disponível em: <http://dx.doi.org/10.1016/0004-3702(90)90060-D>. Citado na página 8.

DORIGO, Marco; BLUM, Christian. Ant colony optimization theory: A survey. **Theoretical Computer Science**, v. 344, n. 2, p. 243 – 278, 2005. ISSN 0304-3975. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0304397505003798>. Citado na página 12.

DORIGO, M.; MANIEZZO, V.; COLORNI, A. Ant system: optimization by a colony of cooperating agents. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 26, n. 1, p. 29–41, Feb 1996. ISSN 1083-4419. Citado na página 12.

DRÉO, Johann; SIARRY, Patrick. A new ant colony algorithm using the heterarchical concept aimed at optimization of multiminima continuous functions. In: ____. **Ant Algorithms: Third International Workshop, ANTS 2002 Brussels, Belgium, September 12–14, 2002 Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 216–221. ISBN 978-3-540-45724-4. Disponível em: <http://dx.doi.org/10.1007/3-540-45724-0_18>. Citado na página 12.

FLORES, M. Julia et al. Handling numeric attributes when comparing bayesian network classifiers: does the discretization method matter? **Applied Intelligence**, v. 34, n. 3, p. 372–385, 2011. ISSN 1573-7497. Disponível em: <http://dx.doi.org/10.1007/s10489-011-0286-z>. Citado na página 8.

FRIEDMAN, Nir; GEIGER, Dan; GOLDSZMIDT, Moises. Bayesian network classifiers. **Machine Learning**, v. 29, n. 2, p. 131–163, 1997. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10.1023/A:1007465528199>. Citado 3 vezes nas páginas 8, 10 e 11.

JIANG, Liangxiao et al. Survey of improving naive bayes for classification. In: ____. **Advanced Data Mining and Applications: Third International Conference, ADMA 2007 Harbin, China, August 6-8, 2007. Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 134–145. ISBN 978-3-540-73871-8. Disponível em: <http://dx.doi.org/10.1007/978-3-540-73871-8_14>. Citado na página 8.

LINDEN, Ricardo. **Algoritmos Genéticos**: Uma importante ferramenta de inteligência computacional. São Paulo: Brasport, 2008. Citado na página 13.

SEBASTIANI, Paola; ABAD, Maria M.; RAMONI, Marco F. Bayesian networks. In: ____. **Data Mining and Knowledge Discovery Handbook**. Boston, MA: Springer US, 2010. p. 175–208. ISBN 978-0-387-09823-4. Disponível em: <http://dx.doi.org/10.1007/978-0-387-09823-4_10>. Citado 2 vezes nas páginas 8 e 10.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. **Data Mining: Practical Machine Learning Tools and Techniques**. 4rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2017. ISBN 0123748569, 9780123748560. Citado na página 8.