Redes Bayesianas para Previsão de Valores de Ações: aplicando os algoritmos K2 e PC

Tulio de Sousa Castro[1], **Edimilson Batista dos Santos**[1]

[1]Universidade Federal de São João del-Rei – UFSJ
Departamento de Computação – DCOMP
Campus Tancredo Neves (CTAN) – São João del-Rei / MG

tuliosousa1998@hotmail.com, edimilson.santos@ufsj.edu.br

Abstract. *Aiming to support the investors' decisions in stock market, in this paper, the application of Bayesian networks for stock market value prediction is proposed. The applied data were obtained through the Yahoo! Finance and they included daily quotations from Petrobras, Telefonica Brasil and Embraer between 02/01/2020 and 11/27/2020. The learning of the Bayesian networks was carried out by the PC and K2 algorithms. For the inference, the exact Lauritzen algorithm was used, which generated good results, with hits of up to 94%, obtaining an average hit rate of 73.66% for the network induced by the PC algorithm and 70.8% for the network induced by K2. Thus, the proposal is promising and manages to satisfactorily forecast the stock price.*

Resumo. *Visando amparar as decisões dos investidores da bolsa de valores, neste trabalho, é proposta a aplicação de redes Bayesianas para predição de valores no mercado de ações. Os dados aplicados foram obtidos por meio do portal Yahoo! Finance e abrangeram cotações diárias das empresas Petrobras, Telefonica Brasil e Embraer entre os dias 01/02/2020 e 27/11/2020. O aprendizado das redes Bayesianas foi realizado pelos algoritmos K2 e PC. Para a inferência, foi utilizado o algoritmo exato de Lauritzen, que gerou bons resultados, com acertos de até 94%, obtendo uma taxa de acerto médio de 73,66% para a rede induzida pelo algoritmo PC e 70,8% para a rede induzida pelo K2. Desta forma, a proposta é promissora e consegue prever satisfatoriamente o preço das ações.*

1. Introdução

Antigamente, para investir em ações, era necessário que as negociações fossem feitas presencialmente e de forma verbal. Tudo era feito em um grande salão onde os operadores executavam as ordens. Entretanto, com os avanços tecnológicos, as negociações passaram a ser completamente digitais e, atualmente, basta criar uma conta em uma corretora e acessar seu *Home-broker*, sistema que conecta o usuário ao pregão eletrônico. A partir dessa facilitação, diversas pessoas foram aderindo ao mercado de capitais e, no Brasil, a cada ano, milhares de pessoas ingressam na bolsa de valores. Com este aumento de investidores, é gerado um enorme fluxo de operações, fomentando assim o surgimento de diversas técnicas para a precificação dos papéis, que vão desde métodos já consolidados até métodos recentemente propostos, como a previsão de valores por meio de Aprendizado de Máquina.

A previsão do mercado de ações é uma tarefa desafiadora, visto que o preço das ações reagem a diversos fatores, como o mercado internacional e acontecimentos políticos. Com os avanços na área de inteligência artificial e na capacidade de processamento de dados, tem surgido diversos métodos na literatura. Alguns trabalhos [Santos 2020] propõem a aplicação de *Support Vector Machine* (SVM) e redes neurais, e têm se mostrado efetivos na previsão de ações. Estes métodos, porém, não permitem a interpretabilidade dos resultados e dificulta assim a tomada de decisão.

Neste artigo, é proposto o uso de um formalismo conhecido como Redes Bayesianas para auxiliar na tomada de decisão no mercado de ações por meio da previsão de valores. Redes Bayesianas têm sido aplicadas em problemas envolvendo incertezas, como é o caso do mercado de ações. Na literatura, é possível encontrar alguns trabalhos que aplicam redes Bayesianas à análise financeira [Malagrino et al. 2018] [Tabassum and Halder 2018], as quais têm se mostrado interessante por proporcionar mais interpretabilidade aos resultados obtidos.

Um dos objetivos deste trabalho é comparar os algoritmos K2 [Cooper and Herskovits 1992] e PC [Spirtes et al. 2000], os quais são bem conhecidos na literatura, para a indução de redes Bayesianas aplicadas à previsão de valores no mercado de ações. Para isto, os algoritmos foram treinados com dados de ações das empresas Petróleo Brasileiro SA Petrobras, Telefônica Brasil e Embraer SA. Os resultados obtidos pelo método de inferência de Lauritzen, aplicado às redes induzidas pelo K2 e PC, mostraram-se satisfatórios e promissores, o que leva a acreditar que a metodologia proposta tende a ajudar investidores na tomada de decisão.

O restante deste texto é organizado como segue. Na Seção 2, é apresentado uma introdução ao mercado de ações. A Seção 3 traz a fundamentação teórica para redes Bayesianas e seu aprendizado. Na Seção 4, são apresentadas a preparação dos conjuntos de dados e aplicação dos algoritmos K2 e PC, assim como os resultados obtidos a partir das redes induzidas por ambos algoritmos. Por fim, a Seção 5 mostra as conclusões obtidas e aponta alguns trabalhos futuros.

2. Mercado de Ações

Uma ação representa a menor parcela de uma empresa e, com o mercado acionário, qualquer pessoa pode se tornar sócio de uma companhia ao comprar uma ação da mesma. Consequentemente, ao se tornar um acionista de uma empresa, além do título, vem a participação em seus resultados. Quando uma empresa de capital aberto lucra, parte de seu lucro é distribuído aos sócios, proporcionalmente à quantidade de ações possuídas.

Diversos métodos para análise de empresas e suas respectivas ações foram desenvolvidos, entretanto, a análise técnica e a análise fundamentalista destacam-se como principais modelos da literatura. A análise técnica baseia-se na identificação de padrões recorrentes a partir de dados passados. Em [Edwards et al. 2012], é afirmado que a análise técnica refere-se ao estudo da ação do próprio mercado em oposição ao estudo das mercadorias em que o mercado negocia.

Dada a simplicidade, ela é amplamente utilizada no Brasil, principalmente pelo pequeno investidor. Diferente da análise fundamentalista, a análise técnica não leva em consideração outros indicadores, como dividendos, gestão, endividamento, lucro, dentre outros. Ela considera o preço passado juntamente com o volume de negociações e

pode ajudar os investidores a acertarem o $timing$ das negociações. A análise técnica também é conhecida como análise gráfica, pois ela é comumente aplicada diretamente sobre gráficos.

Em [Saffi 2003] é dito que uma das principais críticas ao modelo se dá ao fato de que suas conclusões baseiam-se na aplicação das estratégias de análise técnica somente a uma realização do processo estocástico do preço da ação. Uma das formas de contornar este problema seria a utilização de outros modelos que forneçam suporte ao resultado, como a utilização em conjunto da análise fundamentalista ou outros modelos preditivos. A utilização de redes Bayesianas será mais uma alternativa para o amparo das análises de ações feitas por investidores, antes de tomarem suas decisões.

3. Redes Bayesianas

Redes Bayesianas [Marques and Dutra 2002] são grafos acíclicos dirigidos que retratam as dependências entre variáveis de forma probabilística. Segundo [Queiroz 2008], é um modelo gráfico para representar os relacionamentos probabilísticos entre variáveis e realizar inferência probabilística com estas variáveis. De acordo com [Santos 2011], uma rede Bayesiana é formada por dois componentes: a representação gráfica e os parâmetros numéricos (tabelas de probabilidade condicional). Ambos componentes podem ser aprendidos a partir dos dados. A Figura 1 ilustra a estrutura de uma rede Bayesiana.

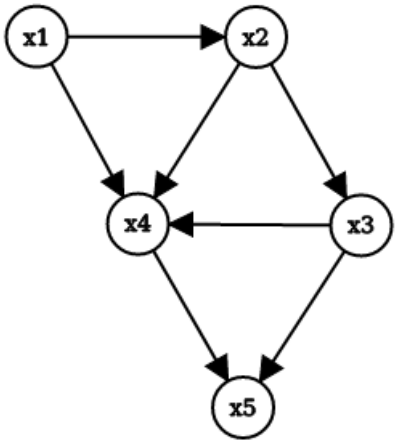

Figura 1. Estrutura de uma rede Bayesiana

Um arco conectando A e B representa um fator na distribuição de probabilidade conjunta, portanto, é necessário saber todos os valores de $P(A|B)$ de forma que seja possível conduzir a inferência. Considerando um conjunto de variáveis aleatórias X = $\{X_1, X_2, ..., X_n\}$, onde cada X_i corresponde a uma variável da rede, é possível definir a probabilidade conjunta de acordo com a equação 1:

$$P(X_1 = x_1, ..., X_k = x_k) = \prod_{i=1}^{k} P[X_i = x_i | pais(X_i)] \quad (1)$$

3.1. Aprendizado de Redes Bayesianas

O processo de aprendizado de redes Bayesianas pode ser dividido em duas etapas: i) o aprendizado numérico, onde é calculada a distribuição de probabilidade conjunta das variáveis e o ii) aprendizado da estrutura, que pode ser visto como o modelo gráfico das dependências. Busca-se a melhor estrutura da rede para a melhor disposição das dependências entre as variáveis.

Os algoritmos de aprendizado da estrutura podem ser divididos em duas principais classes: os que geram a rede por meio de busca heurística e os que utilizam o conceito de independência condicional. A seguir é apresentado o algoritmo K2, pertencente ao grupo de algoritmos que executa uma busca heurística e, em seguida, o algoritmo PC, que faz parte do conjunto de algoritmos baseados em independência condicional. Ambos foram escolhidos para o desenvolvimento deste trabalho por serem bem conhecidos na literatura.

Algoritmo K2 K2 [Cooper and Herskovits 1992] é um algoritmo de busca heurística criado para o aprendizado de estruturas de redes Bayesianas. De acordo com [Souza 2010], o algoritmo recebe como entrada um conjunto de dados e uma ordenação das variáveis e busca entre os $2^{n(n-1)/2}$ tipos de configurações de estruturas de rede qual maximiza a função *score* dada pela equação 2:

$$g(i, r_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{2}$$

onde cada variável $X_i (i = 1, ...n)$ tem r_i possíveis valores $(v_{i_1}, , v_{i_2}, ..., v_{i_{r_i}})$. Cada variável X_i tem um conjunto de pais π_i e q_i é o número de instanciações de π_i. N_{ijk} é o número de objetos no conjunto de dados D, onde X_i tem o valor v_{ik} e π_i é instanciado como w_{ij}, o qual representa a *j-ésima* instanciação relativa a D de π_i. Finalmente, $N_{ij} = \Sigma N_{ijk}$.

Este algoritmo é muito conhecido devido ao seu desempenho em termos de complexidade computacional (tempo) e resultados precisos, obtidos quando uma ordenação de variáveis adequada é fornecida [Santos 2011].

Algoritmo PC A partir de testes estatísticos de independência condicional, o algoritmo PC [Spirtes et al. 2000] busca por uma rede Bayesiana que representa as relações de independência entre as variáveis de um conjunto de dados.

Segundo [Neapolitan et al. 2004], dado um conjunto de independências condicionais em uma distribuição de probabilidade, tenta-se encontrar um grafo acíclico direcionado (DAG), que represente todas e apenas estas independências condicionais de acordo com a condição de *Markov* definida em [Neapolitan et al. 2004].

As principais etapas do algoritmo foram listadas e ordenadas em [Correia 2019]:

1. A partir de um grafo completo, faz a realização de testes de independência e de independência condicional entre cada par de variáveis;
2. Identificação do grafo esqueleto induzido;
3. Identificação das estruturas;

4. Identificação da orientação dos arcos.

A última etapa do algoritmo PC não consegue orientar todas as arestas em alguns casos. Neste caso, seria necessária outra maneira de direcioná-las, como, por exemplo, consultar um especialista.

3.2. Inferência

Inferência é a tarefa de computar a probabilidade de cada variável da rede Bayesiana, quando os outros valores são conhecidos. Segundo [Ehlers 2003], o problema geral da inferência Bayesiana consiste em calcular valores esperados segundo a distribuição a posteriori de θ definida na equação 3:

$$p(\theta_1|x) = \int p(\theta|x)d\theta_2 \tag{3}$$

Os métodos de inferência Bayesiana podem ser divididos em métodos exatos e métodos aproximados. Nos métodos exatos, os valores das probabilidades são calculados de forma exata, sem sacrificar qualquer precisão do resultado, portanto, só são possíveis caso as integrais possam ser calculadas de forma analítica. Já nos métodos aproximados, sacrifica-se precisão em função de velocidade de processamento.

Neste trabalho, a inferência foi realizada através do algoritmo de inferência exata de Lauritzen-Spiegelhalter [Lauritzen and Spiegelhalter 1988]. Este algoritmo efetua um agrupamento de forma a modificar a estrutura da rede em uma estrutura com topologia múltipla. Em [Silveira Júnior 2003], são apresentados mais detalhes.

4. Experimentos e Análise de Resultados

4.1. Preparação dos conjuntos de dados

Neste trabalho, foram escolhidas três empresas listadas na B3 para obtenção de dados de treinamento dos algoritmos K2 e PC: Petrobras, Telefônica Brasil e Embraer.

A primeira ação escolhida foi a PETROLEO BRASILEIRO S.A. PETROBRAS, com símbolo PETR3. Ela está situada na B3 no setor Petróleo, Gás e Biocombustíveis, no segmento Exploração, Refino e Distribuição. É uma das empresas mais importantes de produção de combustível do mundo. As ações da empresa estão presentes no índice IBOVESPA, principal indicador de desempenho das ações negociadas na B3 desde 1968.

A segunda ação escolhida é a TELEFÔNICA BRASIL S.A, com símbolo VIVT3.SA. Esta empresa constitui um dos principais conglomerados de comunicação do mundo, alocada na B3 no setor de Comunicações, com segmento em Telecomunicações.

A terceira e última empresa escolhida para análise do método de previsão de valores foi a EMBRAER SA, de símbolo EMBR3.SA. Atua no setor de Bens Industriais com segmento em Material Aeronáutico e de Defesa, definido pela B3. Alocada em um setor completamente diferente das outras empresas escolhidas.

As cotações diárias históricas destas empresas foram obtidas pelo site da empresa Yahoo! Finance e usadas para criação dos conjuntos de dados que representam cada empresa. As cotações abrangem do dia 01/02/2000 até o dia 27/11/2020, totalizando 5241 cotações por ação. Inicialmente, os conjuntos de dados eram compostos por seis

variáveis: data, abertura, fechamento, fechamento ajustado, maior preço, menor preço e volume. Os atributos 'data' e 'fechamento ajustado' foram excluídos, já que não traziam informações úteis para a inferência. Em seguida, foram excluídos os registros de dias que não ocorreram pregões, como finais de semana e feriados. Os conjuntos de dados também foram discretizados pelo método hierárquico. Como métodos de discretização geram perda de informação no processo, foram realizados experimentos com diversos intervalos, iniciando com 20 intervalos de discretização.

Para o aprendizado das redes Bayesianas, foram utilizados dados que vão do dia 01/02/2000 até o dia 14/07/2014, totalizando 3640 amostras de cada ação. Para a inferência, foram utilizadas novas amostras, compostas por três anos de cotações diárias, a partir do dia 15/07/2014, resultando em 1095 amostras para cada ação. Todos os papéis escolhidos representam diferentes setores da Bolsa, apresentando comportamentos isolados e distintos. Os resultados foram analisados para cotações diárias de cada papel.

4.2. Resultados obtidos a partir do K2 e PC

Os algoritmos K2 e PC foram executados em *Python*, utilizando a biblioteca SMILE[1] (*Structural Modeling, Inference, and Learning Engine*). A Tabela 1 exibe os resultados de inferência obtidos com o algoritmo *Lauritzen* a partir das redes Bayesianas induzidas por K2 e PC. É possível observar que, para a previsão de valor da ação da Petrobras (PETR3), os algoritmos K2 e PC induziram redes que resultaram em 94% de acerto em comparação ao valor real quando se considera o conjunto de dados discretizados em 20 intervalos. Com 30 intervalos, os dois modelos pioram um pouco, obtendo 85% de acerto. A partir de 40 intervalos, os resultados dos modelos ainda se mantêm em queda; com 60 intervalos, a rede do PC resultou em 78% de taxa de acerto e a rede do K2 resultou em 76%.

Analisando a Tabela 1, pode-se notar que os resultados das redes induzidas por K2 e PC para a empresa Telefônica (VIVT3) mantiveram um comportamento parecido com os resultados das redes para PETR3. Para 20 intervalos, ambos modelos obtiveram 83% de acerto; já para 30 intervalos, houve uma queda do desempenho, resultando em 76% acertos com o modelo do PC e 79% para o modelo do K2. Para 40 e 50 intervalos, os modelos apresentaram redução da taxa de acerto, com 77% e 68% para o PC e 69% e 65% para o algoritmo K2. Os resultados continuaram em queda para 60 intervalos. A redução da taxa de acertos é proporcional ao aumento de intervalos.

Para as ações da empresa Embraer (EMBR3), é possível ver, na Tabela 1, que os modelos seguiram a tendência de diminuir a assertividade quando há aumento do número dos intervalos de discretização. Para 20 intervalos, o modelo do PC resultou em 79% de acerto e o modelo do K2 em 77%. Para 30 intervalos, os modelos diminuíram a taxa de acerto, com 76% para o PC e 68% para o K2. Para 50 e 60 intervalos, o modelo do PC continuou apresentando menor desempenho, resultando em 62% e 58% de acertos respectivamente; já o algoritmo K2 resultou em 60% e 56%, respectivamente.

5. Conclusões e Trabalhos Futuros

A partir das evidências de previsibilidade do mercado acionário, foi proposto um método que utiliza redes Bayesianas para auxiliar o investidor na tomada de decisão, por meio

[1] https://www.bayesfusion.com/smile/

Ação	Intervalos	Acertos PC	Confiança PC	Acertos K2	Confiança K2
PETR3	20	94%	89%	94%	88%
PETR3	30	85%	80%	85%	80%
PETR3	40	82%	78%	81%	76%
PETR3	50	71%	75%	78%	73%
PETR3	60	78%	71%	76%	69%
VIVT3	20	83%	78%	83%	77%
VIVT3	30	76%	71%	79%	71%
VIVT3	40	77%	62%	69%	71%
VIVT3	50	68%	59%	65%	61%
VIVT3	60	58%	52%	57%	49%
EMBR3	20	79%	72%	77%	73%
EMBR3	30	76%	68%	68%	75%
EMBR3	40	63%	58%	66%	60%
EMBR3	50	62%	54%	60%	58%
EMBR3	60	53%	47%	56%	53%

Tabela 1. Resultados de inferência obtidos a partir das redes Bayesianas induzidas pelos algoritmos K2 e PC com as bases de dados das empresas Petrobras (PETR3), Telefônica (VIVT3) e Embraer (EMBR3)

.

de um modelo preditivo para valores dos preços das ações. O objetivo deste trabalho foi aplicar, avaliar e comparar o desempenho do algoritmos K2 e PC, como ferramenta no mercado financeiro.

Foram escolhidas três empresas presentes na bolsa de valores brasileiras, são elas, Petróleo Brasileiro SA Petrobras, Telefônica Brasil e Embraer SA. As cotações históricas diárias foram colhidas dos três papéis e utilizadas como base de dados do experimento, resultando em um acerto médio de 73,66% com 67,2% de confiança média para o algoritmo PC e 70,8% de taxa de acerto médio com 67,13% de confiança para o algoritmo K2. As bases de dados foram discretizadas e, como no ato de discretizar os dados há perda de informações, foram feitos experimentos com diversos intervalos de discretização. Quanto maior o número de divisões, maior a fidelidade com os dados originais, já que o intervalo entre os preços é reduzido, se tornando mais fiel aos dados reais. Entretanto, a complexidade do modelo também cresce e o número de acertos tende a diminuir, portanto, há um *trade off* entre resultado e fidelidade no modelo.

Apesar da diminuição da taxa de acertos, o modelo ainda consegue prever tendências de alta e baixa nos preços das ações. Entretanto, VIVT3 e EMBR3 obtiveram valores mais baixos, permanecendo pouco acima de 50% de taxa de acertos. Dentre as três ações, PETR3 se mostrou consideravelmente mais preditiva em relação ao modelo, liderando a taxa de acerto em todos os intervalos de discretização. É importante salientar que apesar da redução nos acertos, o modelo apresentou em todos os casos resultados superiores a 50%, caracterizando um método com possibilidade real de aplicação ao ser utilizado como auxiliar na tomada de decisão no mercado financeiro.

Para trabalhos futuros, é recomendado a exploração de modelos aplicando dados contínuos por apresentarem maior fidedignidade e aplicação real na bolsa de valores. Além disso, é recomendado a criação de uma interface gráfica de forma a facilitar a utilização do modelo pelos investidores.

Referências

Cooper, G. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347.

Correia, J. I. N. (2019). Uma introdução às redes bayesianas. Master's thesis, Universidade da Madeira.

Edwards, R., Bassetti, W., and Magee, J. (2012). *Technical Analysis of Stock Trends*, page 4. Technical Analysis of Stock Trends. Taylor & Francis.

Ehlers, R. S. (2003). Introdução à inferência bayesiana. *URL: http://www. leg. ufpr. br/% 7Epaulojus/CE227/ce227. pdf*.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224.

Malagrino, L. S., Roman, N. T., and Monteiro, A. M. (2018). Forecasting stock market index daily direction: A bayesian network approach. *Expert Systems with Applications*, 105:11–22.

Marques, R. L. and Dutra, I. (2002). Redes bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. *Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil*.

Neapolitan, R. E. et al. (2004). *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ.

Queiroz, C. D. N. (2008). *Redes Bayesianas no gerenciamento e mensuração de riscos operacionais.* PhD thesis, Universidade de São Paulo.

Saffi, P. A. C. (2003). Analise tecnica: sorte ou realidade? *Revista Brasileira de Economia*, 57:953 – 974.

Santos, E. B. (2011). *Aprendizado Indutivo de Redes Bayesianas: Alèm da Precisão na Tarefa de Classificação.* PhD thesis, Universidade Federal do Rio de Janeiro.

Santos, G. C. (2020). Algoritmos de machine learning para previsão da b3. Master's thesis, Universidade Federal de Uberlândia.

Silveira Júnior, L. G. Q. (2003). Uma aplicação de redes bayesianas no auxílio à tomada de decisões médicas. Master's thesis, UFCG.

Souza, A. L. A. (2010). Redes bayesianas: Uma introdução aplicada a credit scoring. *São Carlos, SP*.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press, 2nd edition.

Tabassum, P. and Halder, M. (2018). Stock price forecasting using bayesian network.