Avaliando uma Oportunidade Exploratória de Petróleo através de Mineração de Dados

Marcos A. Affonso, Kate Revoredo, Leila Andrade

Centro de Ciências Exatas e Tecnologia-
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

Avenida Pasteur, 458 - Urca - Rio de Janeiro / RJ - CEP: 22290-240

{marcos.affonso, katerevoredo, leila}@uniriotec.br

Abstract. *A petroleum exploration opportunity (EO) is defined as a mapped region with potential for possessing a sufficient petroleum accumulation that may justify an exploration project. This article proposes to build a predictive model that it is able to economically evaluate a petroleum exploration opportunity through data mining techniques.*

Resumo. *Uma Oportunidade Exploratória de Petróleo (OE) é definida como uma região com potencial de acumulação de petróleo em valor suficiente, que justifique um projeto exploratório. Este artigo propõe a construção de um modelo preditivo capaz de avaliar economicamente esta Oportunidade utilizando técnicas de mineração de dados.*

1. Introdução

Apesar de pesquisas em busca de fontes alternativas de energia como álcool e energia eólica, o consumo de derivados de petróleo no Brasil vem aumentando, como mostra o gráfico de consumo do Ministério de Minas e Energia (Figura 1).

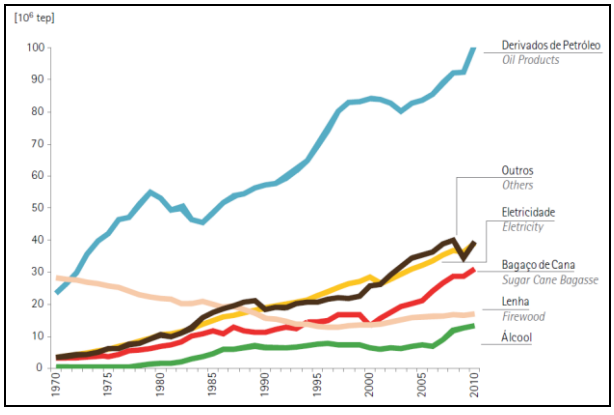


Figura 1. Consumo de energia por fonte no Brasil (Fonte: MME - 2011)

Enquanto fontes alternativas não atingirem um estágio de ampla utilização, a indústria petrolífera vem suprindo a maior parte da energia necessária.

Uma área geográfica que apresenta indícios, ainda não confirmados, de petróleo é denominada Oportunidade Exploratória (OE) e pode dar origem a um projeto que

confirme as expectativas geradas. Pata tal, estudos econômicos são realizados para verificar a viabilidade econômica do projeto.

Neste trabalho, argumentamos que técnicas de mineração de dados podem ser utilizadas para melhorar a avaliação econômica de uma OE através da descoberta automática de um modelo que descreva dados históricos de avaliações econômicas.

O artigo está estruturado da seguinte forma: na Seção 2 explicamos o objetivo de uma avaliação econômica e como ela é feita atualmente. Na Seção 3 descrevemos Mineração de Dados. Na Seção 4 apresentamos a nossa proposta. Na Seção 5 apresentamos os trabalhos relacionados. E, finalmente, na Seção 6 concluímos fazendo as observações finais.

2.Valor Presente Líquido (VPL)

Quando se identifica uma OE promissora, o primeiro passo é elaborar uma curva de produção de óleo, onde a estimativa de produção, ano a ano, do campo é refletida levando-se em conta informações aproximadas sobre a OE.

O segundo passo é o cálculo do Fluxo de Caixa relativo a esta curva de produção. O fluxo de caixa expressa o retorno líquido (receita menos despesa) ano a ano, por toda vida útil do campo.

De posse destas informações, pode-se calcular o VPL que consistirá no somatório dos fluxos de caixa corrigidos para o valor presente. A Figura 2 exibe a fórmula do VPL.

$$VPL = \sum_{k=0}^{n} \frac{(\text{Fluxo\_Caixa})_k}{(1+\text{TMA})^k}$$

Figura 2: Fórmula básica do Valor Presente Líquido (Fonte: Autores)

3. Mineração de Dados

Mineração de dados é parte de um processo maior chamado Descoberta de Conhecimento em Banco de Dados (sigla em inglês KDD) [Han et al 2006]. Frequentemente denomina-se KDD como Mineração de Dados, porém esta é uma das etapas do KDD, onde algoritmos de aprendizado de máquina são aplicados [Mitchell, 1997].

A Figura 3 mostra a sequência de passos de KDD. De início os dados recebem um pré-processamento, como limpeza e tratamento dos dados não observados. Faz parte também dessa etapa a aplicação de técnicas de Normalização e Discretização. A primeira evita que atributos com uma ordem de grandezas maior do que os demais sejam favorecidos, e a segunda visa reduzir o número de valores de um atributo contínuo através da divisão em intervalos. O próximo passo é a mineração de dados, onde são aplicados os algoritmos de aprendizado de máquina para aprender um modelo que reflita os dados.

Por fim, o modelo aprendido passa por um pós-processamento onde os padrões de interesse são filtrados, apresentados de maneira visual e interpretados. O resultado final do processo de KDD é a Informação que levará ao Conhecimento.

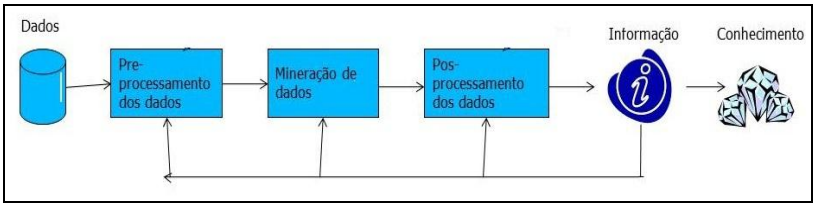


Figura 3: Processo de Descoberta de Conhecimento em Banco de Dados – KDD

Devido ao alto grau de incerteza nas atividades exploratórias, Rede Bayesiana (RB) é o modelo que propomos utilizar para predição do VPL. RB é um modelo que alia Teoria dos Grafos à Teoria da Probabilidade que representa distribuições de probabilidade de modo conciso e utiliza grafos para expressar as dependências entre as variáveis do domínio [Witten et al 2011].

4.Predição do VPL através de RB

A proposta desse artigo é, através de dados históricos de análises econômicas de uma OE, aprender um modelo que seja utilizado tanto para prever o VPL para futuras OEs quanto para descrever o domínio. Para isso o processo de KDD foi aplicado da seguinte forma: (i) Levantamento junto aos especialistas das variáveis relevantes para a análise econômica de uma OE, (ii) Coleta dos dados históricos dessas variáveis, (iii) Aprendizado automático de uma RB a partir dos dados coletados, (iv) Validação da RB aprendida com os especialistas.

Uma estudo de caso foi feito utilizando dados reais de uma empresa de Petróleo, para verificar a viabilidade da proposta. As seguintes variáveis foram consideradas relevantes para o cálculo do VPL: volume de petróleo, lâmina d'agua, profundidade da OE, qualidade do óleo, impostos, TMA (taxa mínima de atratividade), preço do óleo, tipo de fluido (óleo/gás), distância da costa e bacia sedimentar. Com relação às características da rocha onde o Petroleo pode estar depositado, são importantes: área, espessura, porosidade, permeabilidade e saturação. Foram coletados 700 exemplos de avaliações de OE com essas informações.

4.1Estudo de caso: pré-processamento

Os dados foram normalizados entre zero e um ([0;1]). Dessa forma, (i) impedimos que variáveis com grande range numérico dominem as de pequeno range e (ii) ocultamos os verdadeiros valores, compromisso assumido junto à empresa detentora da informação. Além disso, os atributos numéricos foram discretizados. Três métodos foram considerados: (i) bins de igual largura, mas com detecção automática da melhor quantidade de bins para cada variável através do procedimento "leave-one-out", (ii) bins de igual frequência ou (iii) bins de igual largura para todas as variáveis, mesmo que alguns fiquem vazios. A opção (i) mostrou ser a melhor, pois apresentou uma quantidade menor de bins sem representação.

4.2 Estudo de caso: mineração de dados

Após o pré-processamento, a etapa de mineração de dados foi executada, onde algoritmos de aprendizado de redes Bayesianas foram executados. Nós consideramos como algoritmo de aprendizado o K2 local. Além disso, restringimos cada nó da rede a no máximo dois pais. O modelo aprendido apresentou uma acurácia de 70%. Analisando a matriz de confusão resultante, percebeu-se que classes menos representadas apresentam baixas taxas de Verdadeiro Positivo (TP rate).

Uma analise das variáveis mostrou um desbalanceamento do dataset, isto é, algumas classes de VPL possuem pouca representatividade no dataset. Para resolver esse problema, aplicamos o algoritmo SMOTE [Chawla et al, 2002] que tem a finalidade de fazer um resamping aplicando a técnica Synthetic Minority Oversampling. Esta técnica gera exemplos sintéticos das classes menos representadas. Após aplicação do SMOTE houve melhora na acurácia que passou de 70% para 78%. A Figura 4 apresenta a matriz de confusão resultante após aplicação do SMOTE.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as     TP Rate
  13   4   1   2   0   0   0 |   a = '(-inf-0.1]'      0.65
   3 421  45   7   0   2   0 |   b = '(0.1-0.2]'       0.881
   1  44  84   8   1   5   2 |   c = '(0.2-0.3]'       0.579
   0  10  19  36   0   3   4 |   d = '(0.3-0.4]'       0.5
   0   0   1   3  15   0   2 |   e = '(0.4-0.5]'       0.714
   0   0   1   1   2  21   3 |   f = '(0.5-0.6]'       0.75
   0   0   0   0   0   2  37 |   g = '(0.6-inf)'       0.949
```

Figura 4: Matriz de confusão após Resampling com SMOTE

Com o objetivo de encontrar um modelo mais preciso, aplicamos outros algoritmos de aprendizado de RB: Hill Climbing (ao contrário do K2 que assume uma ordem das variáveis, não tem restrição de ordem das variáveis), Tabu (similar a Hill climbing, mas estende um pouco a busca mesmo quando encontra um suposto ponto ótimo). Além disso, testamos algumas opções de parâmetros: (i) Escopo da métrica de avaliação: Local avalia cada nó da rede individualmente e assume que a rede ótima é obtida como o produto das avaliações de todos os nós. Global avalia a rede considerando métricas globais como acurácia, (ii) Número máximo de pais para os nós (2 ou 3).

Usamos como linha base de comparação a rede Bayesiana Naive Bayes. Os resultados obtidos estão na Figura 5, onde cada modelo BN# foi construído com diferentes opções de algoritmo de busca, escopo da métrica e número máximo de pais por nó.

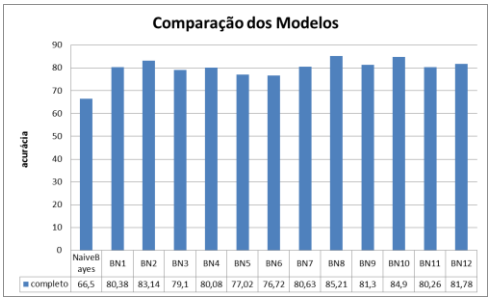
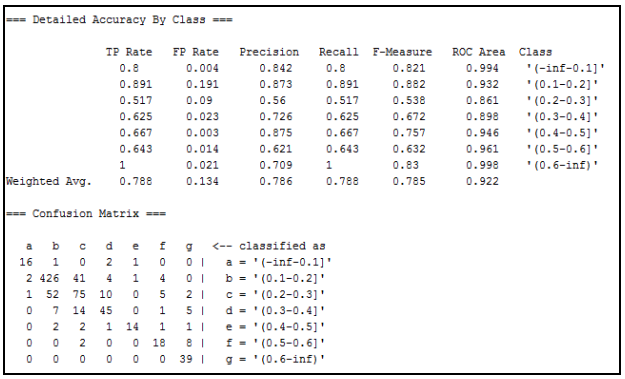


Figura 5: Comparação dos modelos RB

Percebe-se que todas as propostas foram melhores do que o NaiveBayes, sendo o modelo BN8 de melhor desempenho, onde foi utilizado Escopo Global, algoritmo K2 e número de pais igual a 3. Os valores das métricas de avaliação de mineração de dados e a matriz de confusão final para a BN8 são mostradas na Figura 6.



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.004      0.842     0.8       0.821      0.994    '(-inf-0.1]'
                 0.891     0.191      0.873     0.891     0.882      0.932    '(0.1-0.2]'
                 0.517     0.09       0.56      0.517     0.538      0.861    '(0.2-0.3]'
                 0.625     0.023      0.726     0.625     0.672      0.898    '(0.3-0.4]'
                 0.667     0.003      0.875     0.667     0.757      0.946    '(0.4-0.5]'
                 0.643     0.014      0.621     0.643     0.632      0.961    '(0.5-0.6]'
                 1         0.021      0.709     1         0.83       0.998    '(0.6-inf)'
Weighted Avg.    0.788     0.134      0.786     0.788     0.785      0.922

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
  16   1   0   2   1   0   0 |   a = '(-inf-0.1]'
   2 426  41   4   1   4   0 |   b = '(0.1-0.2]'
   1  52  75  10   0   5   2 |   c = '(0.2-0.3]'
   0   7  14  45   0   1   5 |   d = '(0.3-0.4]'
   0   2   2   1  14   1   1 |   e = '(0.4-0.5]'
   0   0   2   0   0  18   8 |   f = '(0.5-0.6]'
   0   0   0   0   0   0  39 |   g = '(0.6-inf)'
```

Figura 6: Medidas de desempenho do modelo BN8

4.3 Estudo de caso: pós-processamento

Após o aprendizado da RB, apresentamos esta aos especialistas para a análise das dependências entre as variáveis ali representadas. Seguindo suas orientações rearranjamos alguns nós e arestas com o objetivo de inserir o conhecimento do domínio e recalculamos as tabelas de probabilidades condicionais.

Construído o modelo, ele foi utilizado para novas inferências a respeito do domínio, como a predição do valor do VPL para uma nova OE. Os resultados obtidos foram satisfatórios.

5.Trabalhos Relacionados

Existem algumas pesquisas no sentido de aplicar a mineração de dados para predição e classificação de valores no domínio da atividade petrolífera. Em [Schoeninger, 2003] é proposto um sistema especialista baseado em Lógica Fuzzy que estima os riscos da exploração petrolífera e calcula a probabilidade de sucesso geológico de uma OE. Schoeninger chegou a fazer experimentos com RB, mas desistiu alegando dificuldades na construção das tabelas de probabilidades condicionais. Neste experimento não houve preocupação com o aspecto econômico da OE, como cálculo do VPL.

Em [Junior, 2010] utiliza-se mineração de dados, especificamente Redes Neurais, para melhorar o gerenciamento da produção de um campo de petróleo por meio da predição da produção período a período. A Rede Neural foi treinada com dados gerados por um simulador e se mostrou uma alternativa eficiente. Nosso trabalho trata de atividade exploratória que é anterior a fase de produção que se caracteriza por ser mais gerencial, e busca a otimização dos procedimentos.

[Martinelli et al, 2011] apresenta uma proposta de Rede Bayesiana que possibilita a análise de um prospecto (grande extensão de área) localizado no Mar do Norte. Explorando as semelhanças geológicas de uma área será possível predizer a viabilidade de um prospecto. Neste estudo, tanto o grafo quanto os parâmetros são definidos pelos especialistas. Não há treinamento da RB como no nosso.

6.Conclusão

Neste artigo expusemos os problemas envolvidos na avaliação de uma Oportunidade Exploratória de petróleo e apresentamos uma proposta para auxiliar essa avaliação através da utilização de mineração de dados.

Nossos testes preliminares apontam na direção de que é possível construir um sistema de apoio a decisão, considerando um modelo preditivo de avaliações econômicas aprendido a partir de dados históricos. Como estudo de caso foram utilizados dados históricos de uma empresa brasileira.

Pretendemos avaliar metodologias como a proposta em [Andrea e Franco,2011], onde uma RB é especificada apartir de uma Ontologia, com o objtivo de encontrar uma rede Bayesiana mais precisa para a tarefa de predição do VPL.

Referências

Andrea, B. e Franco, T. (2011). "Mining Bayesian networks out of ontologies". Journal of Intelligent Information Systems.

Chawla, N; Bowyer, K; Hall, Kegelmeyer, W (2002) "SMOTE: Synthetic Minority Over-sampling Technique", paper, Journal of Artificial Intelligence Research, Morgan Kaufmann Publishers.

Han, J; Kamber, M (2006) "Data Mining: Concepts and Techniques", 2nd ed. Morgan Kaufmann. Pag 5-7.Ben-Gal, I. e Ruggeri, F. e Faltin F. e Kenett R., (2007) "Bayesian Networks" , Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons – Paper

Junior, A. (2010) "Predição não-linear de Curvas de Produção de Petróleo Via Redes Neurais Recursivas", Dissertação de Mestrado – Universidade Federal do Rio Grande do Norte – UFRN – Engenharia de Petróleo

Junior, Repsol (2003) "A Competição e a Cooperação na Exploração e Produção de Petróleo" – COPPE/UFRJ - Dissertação Mestrado. Pag 62-63; 171. Planejamento Energético

Martinelli,G; Eidsvik, J; Hauge, R; Forland, M (2011) "Bayesian Networks for Prospect Analysis in the North Sea", AAPG Bulletin 2011, PP. 1423-1442.

Schoeninger, C. 2003 "Tratamento de Informações Imperfeitas na Análise de Risco de Prospectos em Exploração Petrolífera" - Universidade Federal de Santa Catarina (UFSC) - Dissertação Mestrado em Ciências da Computação/Inteligência Artificial.

Witten, I; Frank, E. (2011) "Data Mining: Practical Machine Learning Tools and Techniques", 3rd ed. Elsevier. Pag 5; 9; 278-279.