

Luiza Torello Vieira
Victor Azadinho Miranda

KDD em Big Data

Motivações

- Dominação da Internet sobre a comunicação global
 - 1993: 1% do tráfego de informação das redes de telecomunicação era feito pela Internet.
 - 2000: 51% do tráfego de informação.
 - 2007: 97% do tráfego de informação.
- O surgimento da Internet aumentou de forma abrupta a quantidade de dados produzidos.
- Internet das coisas fez com que saíssemos da era do Terabyte para o Petabyte.

Motivações

- Novas tecnologias surgiram para extrair informações e padrões de grandes volumes de dados.
- Extração do conhecimento do Big Data é o processo de transformar tais dados em informações acionáveis.
- Com essas informações, análises podem ser feitas.

Desafios

- Transformação de dados em informações não é simples.
- Processo de Mineração de Dados é recheado de desafios.
- Tais desafios podem ser categorizados nos chamados 3V's do Big Data:
 - Volume;
 - Variedade;
 - Velocidade.

Desafios

- Volume:
 - Faz referência ao volume de dados;
 - Constante crescimento.
- Variedade:
 - Múltiplas fontes produtoras;
 - Dados não seguem um padrão.
- Velocidade:
 - Geração de dados extremamente rápida.

Transformação

- Os três Is do KDD em Big Data:
 - Intuição Informada: auxiliar na predição de ocorrências futuras;
 - Inteligência: reconhecer situações atuais para determinar ações;
 - Visão (em inglês: Insight): sabido o que ocorreu, decidir qual ação deve ser tomada.

KDD

- Sigla: **K**nowledge **D**iscovery.
- Processo utilizado para o descobrimento de conhecimento.
- Processo pode ser dividido em 5 etapas:
 - Seleção;
 - Pré-processamento;
 - Transformação;
 - Data Mining;
 - Interpretação.

Etapas do KDD

- Seleção:
 - Definição das fontes de dados.
- Pré-processamento:
 - Aprimoramento do conjunto de dados.
- Transformação:
 - Variáveis, suas importâncias, interações e correlações são analisadas.

Etapas do KDD

- Data Mining (em português: Mineração de Dados):
 - Uso de algoritmos estatísticos e de machine learning;
 - Extração de estruturas, correlações e padrões dos dados.
- Interpretação:
 - Revela caso os padrões detectados são realmente interessantes ou não.

Etapa 1 - Seleção

- Escolher as fontes de dados que se encaixam nas metas do projeto.
- Apenas dados confiáveis devem ser utilizados.
- Certos fatores podem afetar as amostras:
 - Tamanho muito pequeno;
 - Amostras tendenciosas.

Etapa 1 - Seleção

- Definição de qualidade de dados varia.
- Não existe um conjunto universal de características as quais possam descrevê-la.
- Dimensões mais comuns para medir a qualidade:
 - Precisão;
 - Completude;
 - Consistência;
 - Correlação;
 - Atualidade.

Exemplo

- Objetivo: aplicar uma metodologia efetiva para descrever o aparecimento de qualquer problema, e permitir o desenvolvimento de uma ferramenta preditiva nos sistemas de montagem.
- Etapa de seleção:
 - Informações gerais do layout de montagem: códigos de rastreamento e reconhecimento de modelos; data e horário do processamento de itens; tempo total de cada estação de trabalho;
 - Informações de estações específicas: variáveis físicas referentes a operações realizadas; resultados da estação.

Pontuação de Seleção de Dados (em inglês: Data Selection Score)

- Medir a confiabilidade do conjunto de dados
- $D = \{ D_1, D_2, \dots, D_n \}$
- Para cada uma dessas D dimensões, temos os seguintes indicadores de qualidade:
- $Q = \{ q_1, q_2, \dots, q_k \}$

Pontuação de Seleção de Dados (em inglês: Data Selection Score)

- O indicador de qualidade é baseado na razão entre o número de violações V e o número total de instâncias onde a regra é aplicada, multiplicado pelo peso do indicador.
- $Q = (V / T) * \text{peso}_q$

Pontuação de Seleção de Dados (em inglês: Data Selection Score)

- Q é uma razão de erros. A pontuação de medição de qualidade M é calculada através da divisão da soma das pontuações dos pesos dos indicadores de qualidade pelo número total de instâncias nesse contexto, como a fórmula a seguir demonstra.

- $$M = 1 - ((\sum_{i=1}^n Q_i) / | Q_i |)$$

Pontuação de Seleção de Dados (em inglês: Data Selection Score)

- Utilizando essa formulação, definimos a pontuação da seleção de dados como:

- $$\text{PontuaçãoSeleçãodeDados} = \left(\sum_{i=1}^n (\text{peso}_m * M_i) \right) / |M_i|$$

Etapa 2 - Pré-processamento

- Aprimoramento do conjunto de dados.
- Limpeza do conjunto de dados melhora na qualidade do processo.
- Menos ruído nos dados geram resultados mais eficientes.
- Limpeza dos dados:
 - Suplementar valores ausentes;
 - Remover instâncias duplicadas;
 - Resolver inconsistências de dados;
 - Geração de novos atributos.

Etapa 3 - Transformação

- Selecionar um subconjunto ótimo de características que representem os dados e que maximizem a precisão do algoritmo utilizado.
- Principal parte dessa etapa é a seleção de características.

Etapa 3 - Transformação

- A escolha do algoritmo é de extrema importância.
- Escolha incorreta de extração de características pode levar a padrões não representativos.
- Alguns autores consideram como algoritmo confiável aquele que possui uma complexidade de tempo menor.

Etapa 3 - Transformação

- Algoritmo chi-squared.
- Complexidade: $O(n * m)$
 - n: número de amostras
 - m: número de características
- Como o próprio nome sugere, baseado no cálculo do qui-quadrado.
- $$\chi^2 = \frac{N(AN - MP)^2}{PM(N - P)(N - M)}$$

Etapa 4 – Data Mining

- Desafio nesta etapa é a determinação dos métodos e técnicas corretos para utilizar nos dados.
- Para o mesmo método, diversas configurações diferentes podem ser adotadas as quais podem alterar o resultado final.

Etapa 4 – Data Mining

- Fatores que influenciam o desempenho do algoritmo:
 - Sobreajuste: uso de modelos que incluem mais termos que o necessário ou utilizam de abordagens excessivamente complexas;
 - Defeitos do próprio algoritmo: tais como a incompletude de aprendizagem indutiva;
 - Problemas de design do algoritmo: quando o mesmo é facilmente afetado por um item de dados com ruído ou errôneo, detectando um padrão inexistente.

Etapa 5 - Interpretação

- Análise dos resultados.
- Abertura para a avaliação humana.



Obrigado