UNIPAC - CENTRO UNIVERSITÁRIO PRESIDENTE ANTÔNIO CARLOS
CAMPUS BARBACENA

Bacharelado em Ciência da Computação

Mineração de dados

Material de Apoio

Parte III – Análise exploratória dos dados

Prof. Felipe Roncalli de Paula Carneiro
felipecarneiro@unipac.br

2º sem / 2023

Material cedido pela Profª Lívia e Profº Osvano

KDD

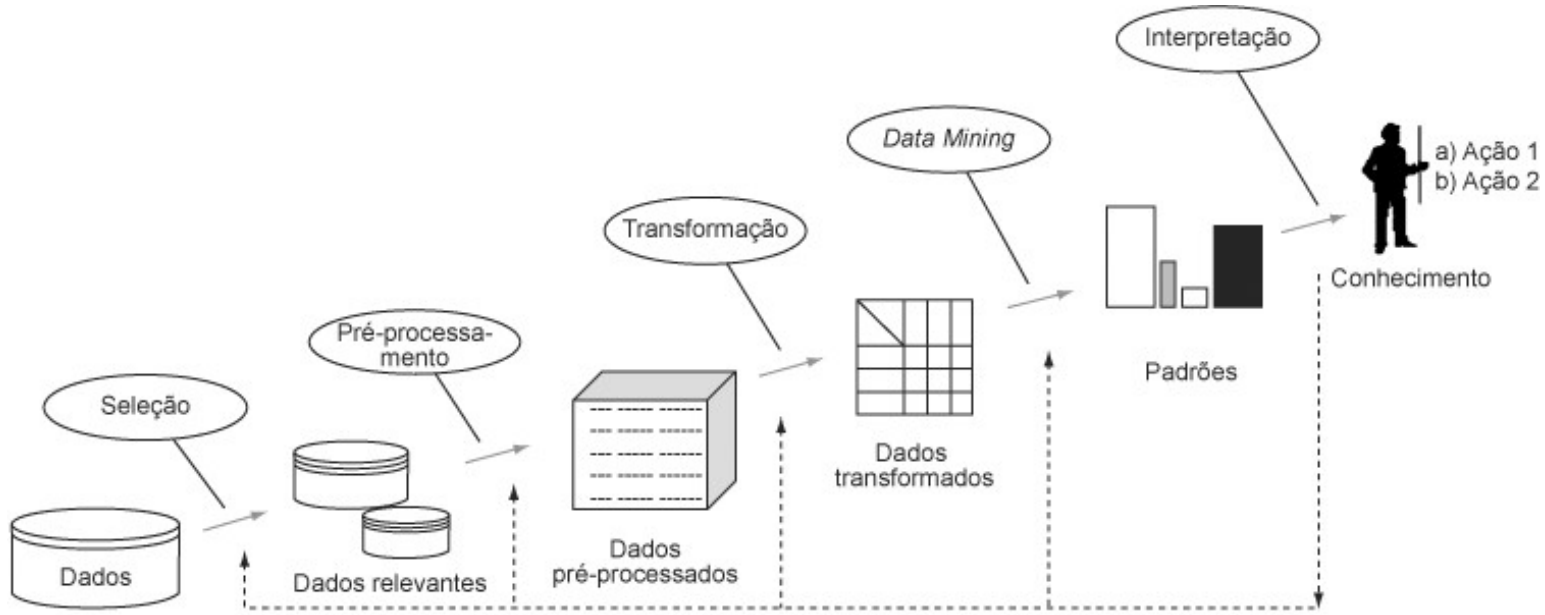

Figura 1. Etapas do processo *KDD* (Fayyad et al. (1996)).

Análise dos dados

- Na fase de mineração de dados, não será possível obter bons resultados se os dois pré-requisitos não forem atendidos:
  - Analista precisa conhecer o contexto em que os dados estão inseridos e como ocorrem nesse contexto;
  - Analista precisa executar procedimentos que tornem o conjunto de dados o mais adequado possível para a etapa de mineração de dados.

Análise dos dados

- Quando a quantidade de dados é muito grande e representativa a tarefa de mineração de dados se faz muito útil e importante.
- Extrapolando a capacidade humana para uma investigação manual, torna-se imprescindível o uso de ferramentas que tenham a capacidade de mostrar diferentes aspectos dos dados, a fim de fornecer “pistas” sobre eles.

Análise dos dados

- Essas “pistas” são interessantes na fundamentação das escolhas referentes aos métodos de análise a serem adotados na fase de mineração de dados.
- Uma das ferramentas mais úteis nesse processo de exploração inicial dos dados é a estatística descritiva.

Análise dos dados

- Aspectos importantes dos conjuntos de dados podem ser obtidos por meio da aplicação de medidas de tendência central, dispersão e correlação, ou também pelo uso de recursos gráficos para visualização dessas e outras medidas.
- Também pode-se perceber através dessa análise as imprecisões e desvios.

Análise dos dados

- Como já observado, esses desvios e outliers influenciam negativamente na mineração.
- A estatística descritiva é útil no planejamento de estratégias de pré-processamento, pois suportam:
  - verificação da presença dos ruídos
  - necessidade de transformação de valores
  - utilidade da seleção de dados ou atributos

Análise dos dados

- Dessa forma, a informação adquirida na análise exploratória apoia a tomada de decisão sobre o pré-processamento e a tarefa de mineração dos dados.

Estatística descritiva

- Uma ferramenta capaz de descrever ou resumir dados, mostrando aspectos importantes do conjunto de dados, como tipo de distribuição associada e os valores mas representativos do conjunto, e permitindo criar visualizações referentes a tais aspectos.

População

- População, ou universo, é o nome que se dá a um conjunto de unidades (elementos ou exemplares) que compartilham características comuns e sobre o qual é pretendido o desenvolvimento de um estudo.
- Exemplo: conjunto de vendas realizadas durante o tempo de funcionamento de um restaurante.

Amostra

- A população pode ser finita ou infinita.
- Se finita, um estudo pode envolver sua totalidade.
- Se muito grande ou infinita, faz-se necessário estabelecer uma *amostra*, para viabilização de um estudo.
- Amostra: subconjunto da população (fração ou parte), que deve ser estabelecido com cuidado, seguindo técnicas apropriadas.

Amostra

- O objetivo de obter uma amostra é reduzir o tamanho da população sem que características essenciais associadas a ela sejam perdidas.
- Boas técnicas de amostragem geram amostras representativas e imparciais da população, ou seja, mantêm a proporcionalidade dos fenômenos que ocorrem na população.

Amostra

- Formas de obtenção de amostra:
  - Aleatória
  - Estratificada
  - Por conglomerados
  - Por quotas
  - E outras

Variável

- Um variável diz respeito a uma característica associada a elementos de uma população.
- É equivalente a um atributo descritivo, podendo ser:
  - Quantitativas (numéricas): podem ser medidas em uma escala de valores numéricos
  - Qualitativas (categóricas): representam uma classificação do elemento ao qual o valor da variável está associado

Medidas de posição e separatrizes

- Medidas de posição (de tendência central) permitem encontrar os valores que orientam a análise dos dados no que diz respeito à sua localização (como a distribuição associada aos valores se comporta no universo amostrado).
- Medidas mais comuns:
  - Média aritmética
  - Mediana
  - Moda

Média aritmética

- Para um conjunto de valores `v = {v1, v2, ..., vn}` em que `n` é quantidade de valores do conjunto, a média aritmética simples é a soma dos valores do conjunto `v`, dividida pela quantidade de valores presentes nesse conjunto:

$$média(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} v_i$$

Mediana

- Mediana é o valor que divide a distribuição dos valores exatamente ao meio (podendo ser vista também como uma medida de separatriz).

- Importante: o valor da mediana não precisa estar presente no conjunto `v`.

Mediana

- Para o cálculo da mediana, todos os valores presentes no conjunto devem ser ordenados de forma crescente (formando o conjunto v') e a mediana será:

$$mediana\,(\mathbf{v}) = \begin{cases} v'_i & \text{se } n \text{ é impar, } i = \dfrac{n+1}{2} \\ \dfrac{1}{2}(v'_i + v'_{i+1}) & \text{se } n \text{ é par, } \quad i = \dfrac{n}{2} \end{cases}$$

Moda

- Moda é o valor mais frequente em um conjunto de valores.

- Pode assumir mais de um valor, quando dois ou mais valores aparecem no conjunto de valores v com a mesma frequência máxima no conjunto.

Exemplo

- Quantidade de vendas, durante um mês, dos nove itens presentes no cardápio promocional de um restaurante.

Item	712	068	002	065	103	809	111	601	044
quant	29	30	32	65	65	65	25	25	90

Exemplo

- Ordenando:

Item	601	111	712	068	002	065	103	809	044
quant	25	25	29	30	32	65	65	65	90

- Análise exploratória dos dados:

Média (quant): 47,3

Mediana: 32

Moda: 65

Exemplo

- A mediana revela que, embora as vendas para o item 002, sejam em quantidade maior que as vendas para metade dos itens observados (é o quinto mais vendido dentro de nove itens), a venda desse item está abaixo da média de vendas do conjunto.

Simetria

- Dessa forma, a comparação dessas medidas pode revelar informações interessantes sobre a distribuição dos valores do conjunto em questão.
- Quando essas medidas assumem o mesmo valor (ou variações muito pequenas entre si) significa que o conjunto de valores de uma variável tem *simetria*, ou seja, mesma distribuição de frequência.

Distribuição assimétrica

- Distribuição assimétrica: se os valores das medidas (média, mediana, moda) são diferentes.
- No exemplo apresentado a distribuição é assimétrica, o que indica que menos da metade dos itens vende mais que a média (quatro itens vendem mais e cinco vendem menos).

Medidas de dispersão

- As medidas de posição (p. ex. média, mediana) são úteis para apresentar uma sumarização dos dados, porém não são capazes de descrever a variação ou dispersão do conjunto de valores.
- Já as medidas de dispersão são capazes de descrever o quanto os valores de um conjunto estão próximos ou distantes de uma medida central.

Medidas de dispersão

- Amplitude: é a diferença entre o maior e o menor valor do conjunto, portanto:

```
amplitude(v) = max(v) - min(v)
```

- Sua análise deve ser feita com cuidado, pois ela pode ser influenciada por valores extremos (outliers).

Medidas de dispersão

- Variância: é uma medida de dispersão definida como a média dos quadrados das diferenças entre cada valor do conjunto `v` e a média desse conjunto.

$$\sigma^2(v) = \frac{1}{n}\sum_{i=1}^{n}(v_i - media(v))^2$$

Medidas de dispersão

- Desvio padrão: é a raiz quadrada da variância. É semelhante à medida de amplitude, com a diferença de que o cálculo do desvio padrão usa todos os valores de um conjunto.
- Geralmente é usado para verificar a consistência de um fenômeno (quando o cálculo do desvio padrão resulta em valores baixos).

$$\sigma\ (v) = \sqrt{\frac{1}{n}\sum_{i=i}^{n}(v_i - media(v))^2}$$

Exemplo

Item	712	068	002	065	103	809	111	601	044
quant	29	30	32	65	65	65	25	25	90

- Amplitude: `max(quant) - min(quant) = 90 - 25 = 65`
- Variância: `513,99`
- Desvio padrão: `22,67`

Referências

- RUSSEL, S., NORVIG, P. *Inteligência Artificial*, Editora Campus, 2ª. edição.
- OLIVEIRA, S. R. M. *Introdução à mineração de dados*, Material para aulas, 2012.
- ZARATE, L.E. *Descoberta de Conhecimento em Banco de Dados e Data Mining*, Material para aulas, 2008.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, MIT, 1996.
- SILVA, L. A. Introdução à mineração de dados com aplicações em R, Elsevier, 1ª ed, 2016.