Pré-processamento de Dados

Danilo Sipoli Sanches

Departamento Acadêmico de Computação
Universidade Tecnológica Federal do Paraná
Cornélio Procópio

Pré-processamento

- Tornar os dados mais adequados para a tarefa de mineração;
– Objetivo: melhorar a mineração com relação a tempo, custo e qualidade;
– Diferentes técnicas podem ser usadas:

- Amostragem
- Discretização ou binarização
- Transformação de variáveis
- Redução de dimensionalidade
- Seleção de atributos

Amostragem

Princípio básico:

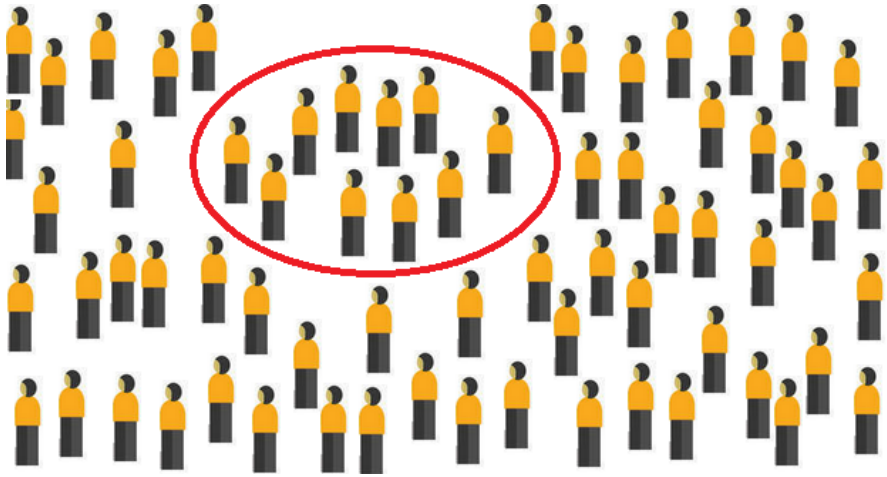
- O uso da amostragem irá produzir resultados tão bons quanto usar o conjunto de dados inteiro, se a amostragem for representativa;
- Menor esforço computacional para processar os dados;

Amostragem representativa

– Possui as mesmas propriedades (de interesse) da base de dados original;
– Ex: mesma média na base original e na amostra

Tipos de Amostragem

- Aleatória;
- Estratificada;
- Progressiva.



Tipo de Amostragem

Amostragem Aleatória sem reposição:

- Cada instância selecionada é removida do conjunto de dados que constituem a população

Amostragem Aleatória com reposição:

- Instâncias não são removidas da população quando ela são selecionadas
- A mesma instância pode ser selecionada mais de uma vez
- A probabilidade de selecionar um objeto se mantêm constante

Tipo de Amostragem

Amostragem Estratificada: Usada para garantir que todas as classes do problema serão representadas

Variações:

- O mesmo número de instâncias de cada classe são selecionadas;
- O número de instâncias selecionadas de cada classe é proporcional ao número de instâncias da classe;

Amostragem Progressiva:

- O tamanho da amostra é difícil de ser determinado;
- Começar com uma pequena amostra e aumentar o tamanho da amostra até que um tamanho suficiente seja encontrado.

Limpeza de Dados

Correção de problemas detectados nos dados deve lidar com:

- Atributos com valores ausentes;
- Atributos e objetos redundantes;
- Atributos e objetos com valores; inconsistentes;
- Atributos com ruídos;
- Outliers.

Exemplo de Valores Ausentes

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
—	M	79	—	38,0	—	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	—	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
—	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Figura: Conjunto de dados com atributos com valores ausentes.

Exemplo de Objetos Inconsistentes

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	**F**	**72**	**Inexistentes**	**38,0**	**3**	**Doente**
19	F	87	Espalhadas	39,0	6	Doente
22	**F**	**72**	**Inexistentes**	**38,0**	**3**	**Saudável**

Figura: Conjunto de dados com objetos inconsistentes.

Exemplo de Objetos Redundantes

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	**F**	**67**	**Inexistentes**	**39,5**	**4**	**Doente**
49	M	92	Espalhadas	38,0	2	Saudável
18	**F**	**67**	**Inexistentes**	**39,5**	**4**	**Doente**
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Figura: Conjunto de dados com objetos redundantes.

Exemplo de Atributos Redundantes

Idade	Sexo	Peso	Manchas	Temp.	# Int.	# Vis.	Diagnóstico
28	M	79	Concentradas	38,0	2	**2**	Doente
18	F	67	Inexistentes	39,5	4	**4**	Doente
49	M	92	Espalhadas	38,0	2	**2**	Saudável
18	M	43	Inexistentes	38,5	8	**8**	Doente
21	F	52	Uniformes	37,6	1	**1**	Saudável
22	F	72	Inexistentes	38,0	3	**3**	Doente
19	F	87	Espalhadas	39,0	6	**6**	Doente
34	M	67	Uniformes	38,4	2	**2**	Saudável

Figura: Conjunto de dados com atributos redundantes.

Tradução de dados

- Ocorre devido a limitações no formato utilizado para armazenar o atributo

- Algumas técnicas podem ter dificuldades com o formato original

- Exemplos:
  - Conversão de hora para valor inteiro;
  - Conversão de data para valor inteiro;
  - Conversão de nome de rua para código postal;

Transformação de Dados

- Mudam o tipo de um atributo;
- Conversão de valores entre tipos;
- Qualitativos para quantitativos (Binarização);
- Quantitativos para qualitativos
- Normalização de valores numéricos
- Tradução de atributos

Normalização

Para normalizar os valores de um atributo:

1. Adicionar ou subtrair uma constante;
2. Multiplicar ou dividir por uma constante;

- Utilizado para mudar intervalo de valores dos dados.

- Permite converter todos os valores de um atributo para o intervalo [0, 1].

Um atributo j de um objeto x_i pode ser calculada como:

$$x_{ij} = \frac{x_{ij} - min_j}{max_j - min_j}$$

sendo min_j e max_j, nessa ordem, os valores mínimo e máximo do atributo j para o conjunto de dados considerado.

Padronização

Para padronizar os valores de um atributo:

1. Adicionar ou subtrair uma medida de localização;
2. Multiplicar ou dividir por uma medida de espalhamento;

Se os valores têm uma distribuição Gaussiana

- Subtrair a média
- Dividir pelo desvio padrão
- Produz valores com distribuição normal (0,1)

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Maldição da dimensionalidade

Número de possíveis objetos cresce exponencialmente com aumento do número de atributos preditivos:

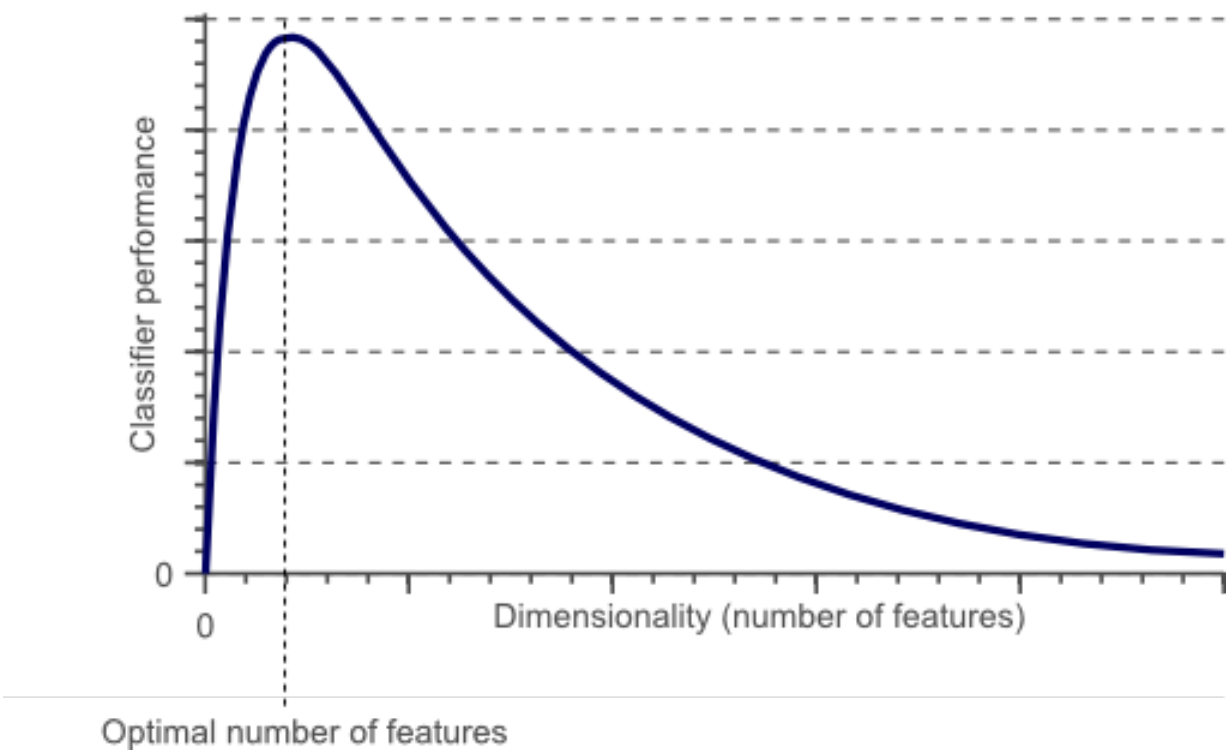
- Espaço formado por 1 atributo com 10 possíveis valores: 10 possíveis objetos;
- Espaço formado por 5 atributos com 10 possíveis valores: 10^5 possíveis objetos; Problemas com poucos objetos e muitos atributos: Dados se tornam muito esparsos.

Maldição da dimensionalidade

Dados esparsos:

- Sem objetos em várias das regiões do espaço de objetos;
- Distâncias entre objetos convergem para um mesmo valor;
- Objetos tendem a se tornar equidistantes;
- Prejudica desempenho de algoritmos que usam distância para definir similaridade entre objetos.

Maldição da dimensionalidade



Maldição da dimensionalidade

- Número de objetos necessários para que modelo induzido tenha um bom desempenho preditivo;
- Cresce exponencialmente com o número de atributos preditivos;
- Na prática, número de objetos em um conjunto de dados é fixo
  Alternativa: redução de dimensionalidade.

Redução da dimensionalidade

- Alguns conjuntos podem ter um número muito grande de atributos;

  Ex1.: Um atributo preditivo para frequência de cada palavra que aparece em um texto;

- Reduzir dimensão;
- Agregação (construção) de atributos;
- Criar novos atributos combinando atributos originais;
- Seleção de atributos;

Seleção de atributos

- Identificar atributos importantes;
- Melhorar desempenho de algoritmo de para indução de modelos;
- Minimizar os efeitos de ruídos;
- Reduzir custo de coleta de dados.

Abordagem Filtro

- Seleção de atributos independe do algoritmo de AM utilizado;
- Verifica co-relação entre atributos;
- Conseguem lidar de forma eficiente com uma grande quantidade de objetos e de atributos preditivos;
- Baixo Custo Computacional;

Abordagem Filtro

Desvantagem:

- Não interage com o algoritmo de indução de modelos;
- Não levar o viés do algoritmo em consideração pode levar a modelos poucos eficientes.

Abordagens *Wrappers*

- Utilizam algoritmo de indução de modelos para guiar seleção de atributos preditivos;
- Ex. Atributos que levaram a menos erros de classificação para um algoritmo de AM;
- Melhor conjunto para um dado algoritmo de indução de modelos;
- Geralmente melhora desempenho obtido pelos modelos induzidos.

Abordagens *Wrappers*

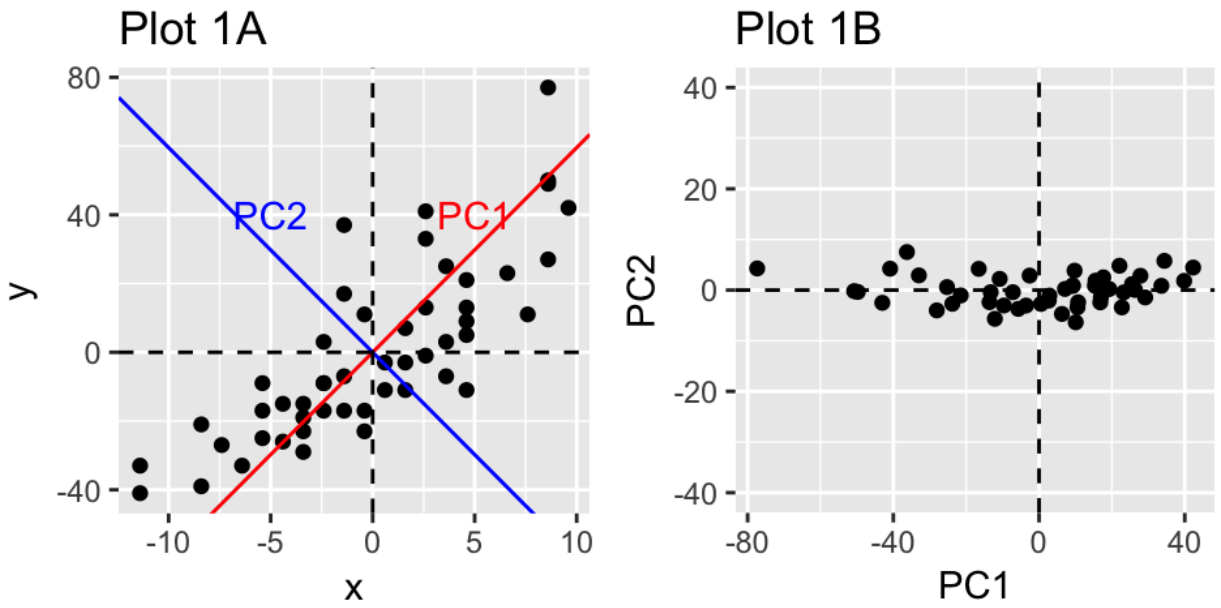
Desvantagens:

- Risco de *overfitting*;
- Desempenho depende do algoritmo de indução;
- Custo computacional elevado, por causa do grande número de execuções do algoritmo de indução;
- Precisa ser repetido quando um novo algoritmo de indução for utilizado.

Análise de Componentes Principais

- O método de Análise de Componentes Principais (PCA) é um método linear de extração de atributos;
- Baseia-se em uma transformação linear sobre os dados;
- Obtêm projeções sobre um novo sistema de coordenadas ortogonais;
- Permite representar a maior variância possível dos dados, numa sequência decrescente (componentes);

Análise de Componentes Principais



(a) Dados em sistema de coordenadas X e Y.

(b) Redução da dimensionalidade. Projeção das amostras para o eixo PC1.

Figura: Exemplo PCA.

Referências

- Faceli K, Lorena AC, Gama J, Carvalho ACP de LF de. Inteligência artificial: uma abordagem de aprendizado de máquina. 2011;
- Jolliffe Ian T. and Cadima Jorge 2016 - Principal component analysis: a review and recent developments;
- https://medium.com/@raevskymichail/feature-selection-overview-of-everything-you-need-to-know-598c53c01d46;
- http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/;