TRANSFORMAÇÃO DE DADOS
REDUÇÃO DE DIMENSIONALIDADE

Cristiane Neri Nobre

Redução de dimensionalidade

- Muitos problemas que podem ser tratados por técnicas de AM apresentam um número elevado de atributos.
- Poucas técnicas de AM podem lidar com um número tão grande de atributos
- O efeito do número muito grande de atributos em algoritmos de AM é descrito como problema da **maldição de dimensionalidade**

Redução de dimensionalidade

- Suponha conjunto de dados em que cada instância possui apenas **um atributo** e que esse atributo pode assumir um dentre **10 valores**

- Esse conjunto de dados pode ter então 10^1 ou 10 instâncias diferentes, um para cada valor diferente do atributo

- Se o número de atributos passar para 5, o número de possíveis instâncias passa a ser 10^5, que é um número de possíveis **instâncias muito maior** do que quando apenas um atributo foi utilizado.

Redução de dimensionalidade

- Uma forma de minimizar o impacto do problema da dimensionalidade é **combinar ou eliminar** parte dos atributos irrelevantes

- A redução do número de atributos pode ainda melhorar o **desempenho** do modelo induzido, reduzir seu **custo computacional** e tornar os resultados obtidos mais **compreensíveis**

Redução de dimensionalidade

As abordagens utilizadas para resolver este problema podem ser divididas em:

- **Agregação**
  - Substituem os atributos originais por novos atributos formados pela combinação de grupos de atributos
  - Levam à perda dos valores originais dos atributos, o que pode ser importante dependendo do contexto (finanças, saúde, etc)

- **Seleção de atributos**
  - Mantem uma parte dos atributos originais e descartam os demais atributos

Redução de dimensionalidade

- **Seleção de atributos** permite:

- Identificar atributos importantes
- Melhorar o desempenho de várias técnicas de AM
- Reduzir a necessidade de memória e tempo de processamento
- Eliminar atributos irrelevantes e reduzir ruído
- Lidar com a maldição da dimensionalidade
- Simplificar o modelo gerado e tornar mais fácil a sua compreensão
- Facilitar a visualização dos dados
- Reduzir o custo de coleta de dados e com isso aumentar acesso a novas tecnologias

Redução de dimensionalidade

Três abordagens são utilizadas para avaliar a qualidade ou desempenho de um subconjunto de atributos:

- **Embutida**
- **Baseada em Filtro**
- **Baseada em Wrapper**

Redução de dimensionalidade

- **Embutida**
  - Nesta abordagem a seleção do subconjunto é embutida ou integrada no próprio algoritmo de aprendizado (ex: Árvore de decisão)

- **Baseada em Filtro**
  - Nesta abordagem, em uma etapa de pré-processamento, é utilizado um filtro sobre o conjunto de atributos original que filtra um subconjunto de atributos do conjunto original, sem levar em consideração o algoritmo de aprendizado que utilizará esse subconjunto (Ex: correlação)

- **Baseada em Wrapper**
  - Utiliza o próprio algoritmo de aprendizado como uma caixa-preta para a seleção;
  - Geralmente é utilizado junto com uma técnica de amostragem;
  - Para cada possível subconjunto, o algoritmo é consultado e o subconjunto que apresentar a melhor combinação entre redução da taxa de erro e redução do número de atributos é em geral selecionado

Redução de dimensionalidade

Vantagens das abordagens baseadas em Filtro:

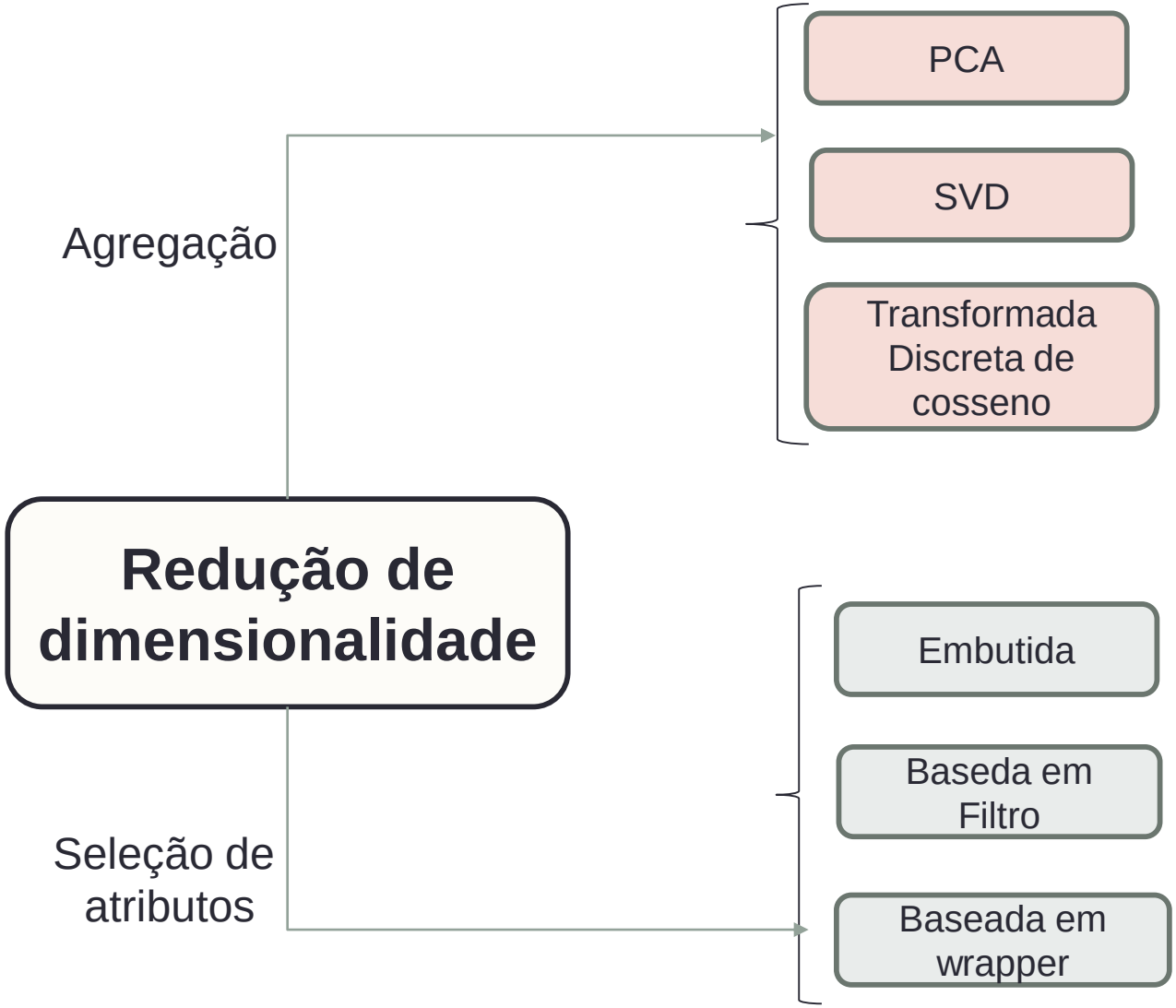
- Como o processo de seleção não depende de nenhum indutor, as características selecionadas podem ser utilizadas por diferentes algoritmos de AM;
- As heurísticas utilizadas para avaliar um subconjunto são computacionalmente pouco custosas, assim os filtros podem ser bastante rápidos
- Os filtros conseguem lidar eficientemente com uma grande quantidade de dados

Redução de dimensionalidade

Vantagens das abordagens baseadas em Wrapper:

- Podem ser muito eficientes (quanto às métricas de avaliação), apesar do tempo computacional
- É uma alternativa simples

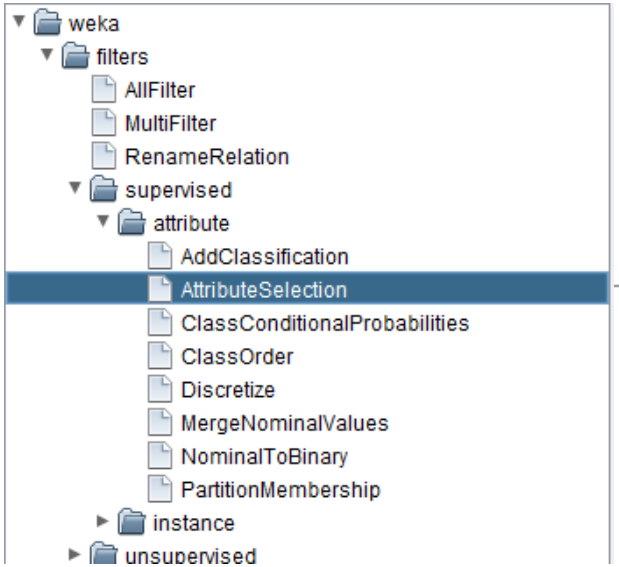
Redução de dimensionalidade



Redução de dimensionalidade

Como realizar a **seleção dos atributos** no ambiente WEKA?

- Carregue o arquivo: weather.nominal.arff que fica na pasta **Data** onde o WEKA está instalado
- Na tela principal do WEKA, vá até a opção **weka/Filters/supervised/atribute/attributeSelection**



Um filtro de atributo supervisionado que pode ser usado para selecionar atributos

Redução de dimensionalidade

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Base Original

Atributos selecionados

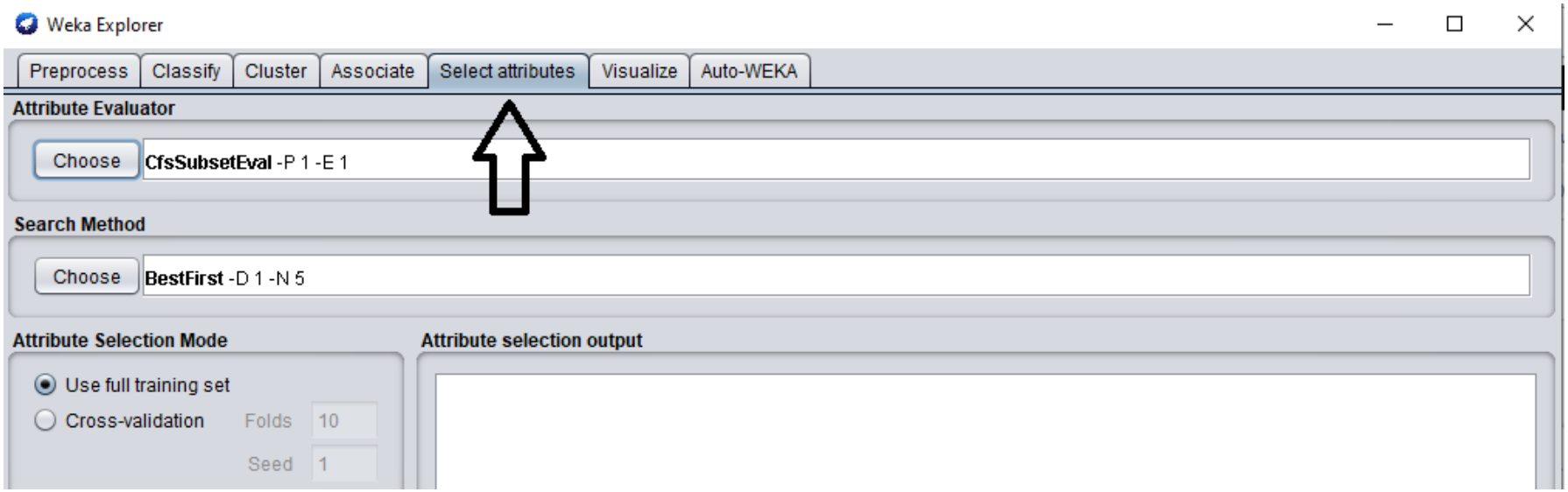
No.	1: outlook	2: humidity	3: **play**
	Nominal	Nominal	Nominal
1	sunny	high	no
2	sunny	high	no
3	overcast	high	yes
4	rainy	high	yes
5	rainy	normal	yes
6	rainy	normal	no
7	overcast	normal	yes
8	sunny	high	no
9	sunny	normal	yes
10	rainy	normal	yes
11	sunny	normal	yes
12	overcast	high	yes
13	overcast	normal	yes
14	rainy	high	no

Neste caso, o modelo pode ter um resultado pior!
Importante testar!

Redução de dimensionalidade

Além disso, o WEKA tem uma aba específica para seleção de atributos!

Investiguem!



Referências:

Capítulo 3 do livro (Seção 3.7)

- Katti Faceli et al. Inteligência Artificial, Uma abordagem de Aprendizado de Máquina, LTC, 2015.