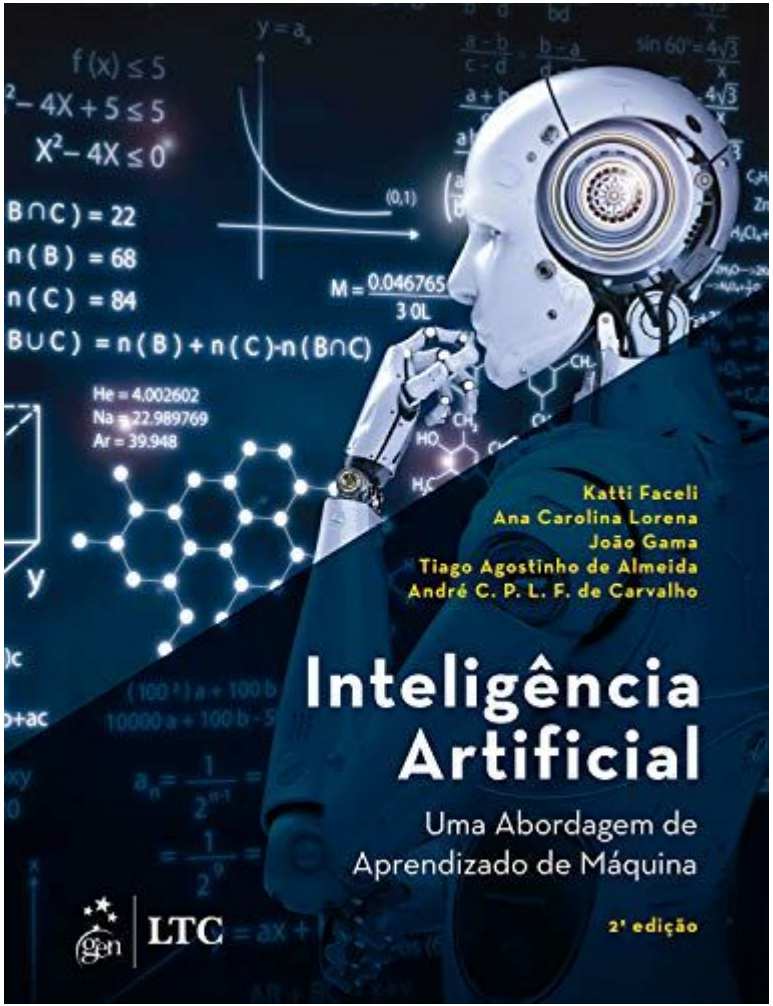
Tratamento de Dados

Integração, Limpeza e Transformação

Aula baseada no Capítulo 3 do livro:



Integração

Integração de Dados

- **Problema de Identificação de entidade**: identificação dos objetos presentes em diferentes conjuntos de dados a serem integrados.
- **Solução**: encontrar atributos comuns aos objetos nos conjuntos de dados a serem integrados.
- **Exemplo**: Considere vários conjuntos de dados médicos. Busca-se pelo valor do Id de um paciente (atributo) que aparece nos em objetos dos diferentes conjuntos. Logo, objetos com aquele Id se referem a um determinado paciente e podem ser combinados em um único objeto do conjunto integrado.

Integração de Dados

- Nem tudo é perfeito:
  - Atributos equivalentes podem aparecer com nomes diferentes em conjuntos de dados distintas.
  - Dados a serem integrados podem estar desatualizado (atualizados em momentos diferentes).
- **Metadados**: dados sobre dados
  - Descrevem as principais características dos dados, evitando erros na integração.

Integração de Dados

- **Data warehouse**: repositório que centraliza informações para análise e tomada de decisões.
- Diferentes sistemas, bancos de dados e outras fontes de dados enviam dados para a data warehouse.

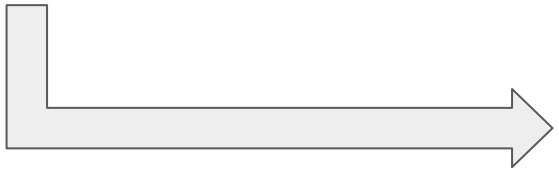
Integração de Dados

- Um número elevado de dados pode prejudicar o **desempenho** de algoritmos de aprendizado de máquina e, muitas vezes, torna-se mais eficiente usar **parte do conjunto de dados**.
- **Maldição da dimensionalidade**: o aumento na dimensão dos dados, ou seja, no **número de atributos** gera um aumento **exponencial** na demanda por dados para treinar **modelos** de aprendizado de máquina **gerais** e **precisos** o suficiente.
- Os objetos se tornam **esparsos** e **equidistantes**.
    - Por exemplo, **clusterizações** não podem ser formados se as distâncias forem equidistantes entre os dados, uma vez que técnicas de agrupamento usam a distância euclidiana para avaliar similaridades.

Remoção de Atributos

- A **experiência** daqueles que conhecem o domínio de dados costuma ser fundamentais na **seleção** de quais **atributos** devem formar o conjunto de dados para se conduzir determinada análise.

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável



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
34	M	67	Uniformes	38,4	2	Saudável

Amostra de Dados

- **Aumento no número de dados** pode gerar modelos com **maior acurácia** ao custo de **eficiência computacional**.
- **Amostra ou subconjunto de dados**: permite alcançar um melhor compromisso entre **acurácia** e **eficiência computacional**.
- Modelos gerados com amostra de dados podem apresentar uma acurácia similar aos gerados utilizando o conjunto completo de dados.
- Uma **amostra representativa** do conjunto de dados deve ser definida, caso contrário, não representará o problema a ser modelado.

Amostra de Dados

- Os dados da amostra devem apresentar a mesma distribuição estatísticas do conjunto de dados original.
- Não há garantias de que a amostra reflita a distribuição do conjunto original, mas algumas técnicas podem ser aplicadas na definição da amostra de dados:
  - Amostragem aleatória simples.
  - Amostragem estratificada.
  - Amostragem progressiva.

Amostra de Dados

- **Amostragem aleatória simples**:
  - **Sem reposição** dos objetos selecionados do conjunto original de dados para compor a amostra.
  - **Com reposição** dos objetos selecionados, ou seja, uma cópia do objeto da conjunto original de dados é incluída na amostra.
- A probabilidade de seleção fica constante com a reposição dos objetos.
- Não há diferença relevante quando as duas técnicas são aplicadas para gerar uma **amostra bastante reduzida** do conjunto original de dados.

Amostra de Dados

- **Amostragem estratificada**: aplica-se quando há **classes com propriedades** diferentes como, por exemplo, quantidade de objetos em cada classe.
  - Amostrar a mesma quantidade de objetos para cada classe.
  - Amostrar a quantidade de objetos para cada classe proporcionalmente ao conjunto original.
- **Problema**: A distribuição dos dados pode impactar a formação das classe em problemas de classificação.
  - **Classes majoritárias** da amostra podem levar a modelos de classificação tendenciosos.
  - Temos o problema de **dados desbalanceados**.

Amostra de Dados

- **Amostragem progressiva**: define-se primeiro uma amostra reduzida que aumenta à medida que a inserção de novos dados na amostra melhoram a acurácia.
- Permite definir a **quantidade mínima de dados** na amostra com reduzida perda de acurácia.
- Uma nova amostra com a mesma quantidade mínima de dados pode confirmar a acurácia do tamanho encontrado.

Automação da Amostra de Dados

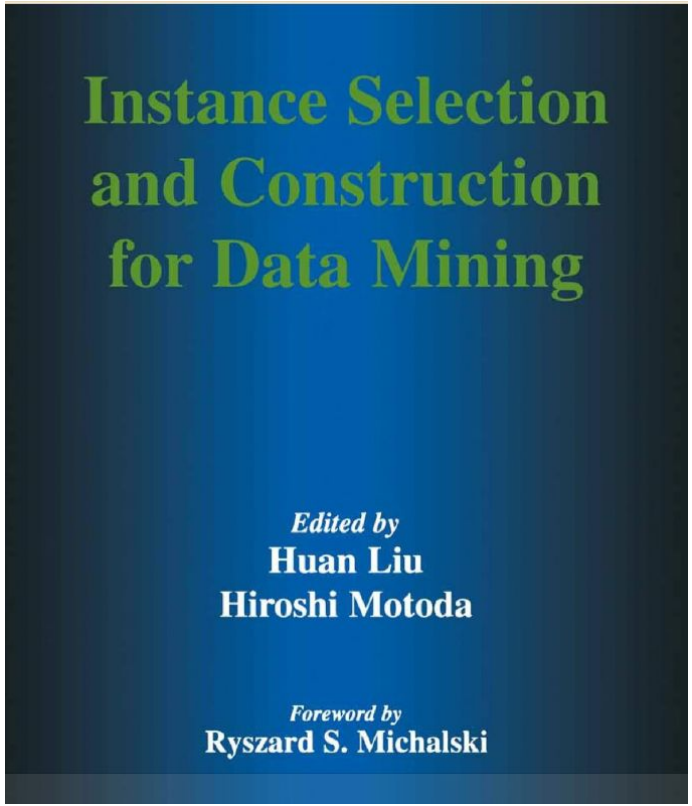
Home > Data Mining and Knowledge Discovery > Article

Published: April 2002

Advances in Instance Selection for Instance-Based Learning Algorithms

Henry Brighton & Chris Mellish

Data Mining and Knowledge Discovery **6**, 153–172 (2002) | Cite this article



Dados Desbalanceados

- Dados desbalanceados surgem em problemas de classificação de dados.
  - 80% pacientes doentes (classe majoritária) e 20% saudáveis (classe minoritária)
  - Tendência do modelo a classificar na classe majoritária.
- Critério aceitável: a acurácia preditiva para um conjunto de dados desbalanceados precisa ser maior que a acurácia obtida atribuindo todo novo objeto à classe majoritária.

Dados Desbalanceados

- **Possível solução**: gerar novos dados seguindo a mesma distribuição estatística ou procedimento que gerou o conjunto de dados atual.
- **Inviável na prática!!**
- Técnicas propostas seguem em geral três abordagens:
  - Redefinir o tamanho do conjunto de dados.
  - Utilizar diferentes custos de classificação para as diferentes classes.
  - Induzir um modelo para uma classe.

Dados Desbalanceados

- **Redefinir o tamanho do conjunto de dados** pela remoção de objetos da classe majoritária ou acréscimo de objetos à classe minoritária
- Acréscimo de novos objetos não pode:
  - levar a cenários ou **situações que nunca ocorrerão na prática**, prejudicando o aprendizado do modelo;
  - levar ao **overfitting**, ou seja, modelo extremamente ajustado aos dados de treinamento.
- Eliminação de objetos pode:
  - levar a **retirada de dados relevantes** para o aprendizado do modelo;
  - ocasionar o **underfitting**, onde o modelo não se ajusta aos dados do treinamento.

Dados Desbalanceados

- **Utilizar diferentes custos de classificação para as diferentes classes** apresenta como desafio a definição adequada de tais custos.
  - Por exemplo, se a **classe majoritária tem o dobro de objetos da classe minoritária**, quando o modelo classifica incorretamente um objeto da classe minoritária, isso equivale à ocorrência de **dois erros de classificação para um objeto da classe majoritária**. Entretanto, a definição dos diferentes custos geralmente não é tão direta
- **Induzir um modelo para uma classe**, ou seja, utiliza-se apenas um modelo de classificação para cada classe.
  - A classe minoritária, a classe majoritária, ou ambas são aprendidas separadamente.
  - Os modelos aprendem utilizando apenas objetos da classe positiva (classe tratada).

Limpeza

Limpeza de Dados

- Busca tratar situações que afetam a qualidade dos dados.
- Situações causadas em geral por
  - problemas em equipamentos ou sensores que realizam a coleta, transmissão e armazenamento dos dados;
  - problemas no preenchimento ou na entrada dos dados por seres humanos.
- Se não tratados, podem influenciar na performance de métodos ou na análise dos resultados obtidos.
- Causas que levam à limpeza de dados:
  - Dados ruidosos
  - Dados inconsistentes
  - Dados redundantes
  - Dados incompletos

Limpeza de Dados

- **Dados ruidosos:** possuem erros ou valores que são diferentes do esperado.
- **Inconsistentes:** não combinam ou contradizem valores de outros atributos do mesmo objeto.
- **Redundantes:** dois ou mais objetos têm os mesmos valores para todos os atributos ou dois ou mais atributos têm os mesmos valores para todos os objetos.
- **Incompletos:** ausência de valores para alguns dos atributos em parte dos dados.

Dados Incompletos

- **Dados incompletos:** ausência de valores para alguns dos atributos em parte dos dados.
- **Eliminar os objetos com valores ausentes.**
  - Recomendada quando um dos atributos faltantes no objeto é o que indica a sua classe.
  - Não recomendada quando
    - Há poucos atributos faltantes no do objeto.
    - O número de atributos com valores ausentes varia muito entre os objetos.
    - O número de objetos que restarem for pequeno.

Dados Incompletos

- **Definir e preencher manualmente valores para os atributos com valores ausentes.**
  - Não pode ser utilizada quando o número de objetos ou atributos com valores ausentes for muito grande.
- **Aplicar um método ou heurística para automaticamente definir valores ausente para atributos.**
  - Alternativa mais utilizada.
- **Utilizar técnicas de aprendizado de máquina que lidam com valores ausentes.**
  - Por exemplo, algoritmos indutores de árvores de decisão.

Dados Incompletos

- Primeira abordagem para **definição automática de valores ausentes**:
  - **Utilizar um novo valor para indicar que aquele atributo tinha um valor ausente.**
    - O novo valor pode ser o mesmo para todos os atributos, ou um valor diferente para cada atributo.
    - O algoritmo aplicado pode entender que tal valor indica algo relevante para o problema.

Dados Incompletos

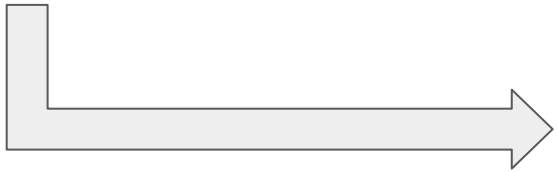
- Segunda abordagem para definição automática de valores ausentes:
  - **Utilizar a média, moda (no caso de valor simbólico) ou mediana dos valores conhecidos para esse atributo.**
    - Calcula-se a métrica para todos os objetos ou apenas os objetos da mesma classe cujo objeto apresenta valor ausente para o atributo.
    - Utilizar o valor mais frequente nos k objetos mais semelhantes ao objeto com valor faltante.
    - Se os objetos apresentam relação no tempo, pode-se calcular a medida considerando objetos nos instantes imediatamente anterior e posterior ao objeto com atributo faltante.

Dados Incompletos

- Terceira abordagem para definição automática de valores ausentes:
  - **Utilizar um indutor para estimar o valor do atributo.**
    - Aplica-se quando o **atributo alvo está ausente**.
    - Os demais atributos são utilizados como atributos de entrada para inferir o valor do atributo ausente.
    - Toma-se como base o valor empregado em objetos semelhantes.

Dados Incompletos

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



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

Dados Inconsistentes

- Apresentam valores conflitantes em seus atributos.
- Inconsistência entre valores de atributos de entrada
    - Valor 520 para o atributo Peso e o valor 1 para o atributo Idade
- Inconsistência entre os valores dos atributos de entrada e o valor do atributo alvo
    - Dois pacientes com exatamente os mesmos valores para os atributos de entrada e diagnósticos diferentes.
- Inconsistências nas relações conhecidas entre os atributos.
    - Sabe-se que os valores de um atributo variam de forma inversamente proporcional em relação aos valores de um outro atributo

Dados 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**

Dados Redundantes

- Há tanto objetos quanto atributos redundantes.
- **Objeto redundante** é aquele com valores dos atributos muito semelhante a outro objeto no mesmo conjunto de dados.
- A técnica de aprendizado de máquina pode dar ao objeto repetido uma relevância maior.
    - Um objeto A com duas cópias pode ser considerado três vezes mais importante que um objeto B não duplicado.
- Há exceções, por exemplo, Boosting é uma técnica de aprendizado de máquina que duplica a quantidade de objetos difíceis de ser classificada.

Dados 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

Dados Redundantes

- **Atributo redundante** apresenta valores para os objetos que podem ser obtidos a partir do valor de outros atributos.
    - Um atributo como **Idade** e **Data de Nascimento** no mesmo conjunto de dados.
- Atributo redundante também aparece através da sua correlação com outros atributos do conjunto de dados.
- Atributos correlacionados apresentam um perfil de variação semelhante para os diferentes objetos.
- Por outro lado, se tal correlação aparece entre um atributo de entrada e um atributo alvo, isso significa uma grande influência do atributo de entrada na predição do valor do atributo alvo.

Dados 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

Dados com Ruídos

- **Dados com ruídos** apresentam objetos que parecem **não pertencer à distribuição** que gerou os dados analisados.
- Ruído pode ser definido como uma **variância ou erro aleatório** no valor gerado ou medido para um atributo (Han e Kamber, 2000).
- Ruídos podem gerar **dados inconsistentes**.
- **Outliers** podem ser um indício da presença de ruído.
- **Outliers** são valores que ultrapassam os limites aceitáveis, ou são muito diferentes dos demais valores observados para o mesmo atributo indicando exceções **raramente** vistas.

Dados Redundantes

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	300	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
34	M	67	Uniformes	38,4	2	Saudável

Dados com Ruídos

- Abordagens para remoção de ruído:
  - Técnicas de encestamento
  - Técnicas baseadas em agrupamento dos dados
  - Técnicas baseadas em distância
  - Técnicas baseadas em regressão ou classificação

Dados com Ruídos

- **Técnicas de encestamento:**
  - Os valores encontrados para esse atributo em todos os objetos são **ordenados**.
  - Os valores ordenados são divididos em faixas ou cestas com o **mesmo número de valores**.
  - Os valores em uma mesma cesta são substituídos, por exemplo, pela média ou mediana dos valores presentes na cesta.

Dados com Ruídos

- **Técnicas baseadas em agrupamento dos dados:** aplicadas tanto para **objetos quanto para atributos.**
  - Para atributos, seus valores são agrupados usando uma técnica de agrupamento.
  - Valores de atributos fora dos grupos são considerados ruídos ou outliers.
  - Para objetos, também são agrupados e o objeto ruidoso é identificado ao ser colocado em um grupo no qual os demais objetos pertencem a uma outra classe.

Dados com Ruídos

- **Técnicas baseadas em distância:** verificam a que classe pertencem os objetos mais próximos de cada objeto **x**.
    - Se os objetos mais próximos pertencem a uma outra classe, são boas as chances de x apresentar ruído, embora possa apenas estar próximo à fronteira de separação das classes.
    - Objetos nesta situação podem ser considerados relativamente instáveis, uma vez que mesmo uma pequena quantidade de ruídos pode movê-los para o lado incorreto da fronteira.

Dados com Ruídos

- **Técnicas baseadas em regressão ou classificação:** aplicam uma função de regressão para, dado um valor com ruído, estimar seu valor verdadeiro.
    - Se o valor a ser estimado for simbólico, uma técnica de classificação pode ser utilizada.

Transformação

Transformação de Dados

- Técnicas de aprendizado de máquinas podem estar limitadas à manipulação de valores de determinados tipos, por exemplo, apenas **valores numéricos** ou apenas **valores simbólicos**.
- Algumas técnicas têm seu desempenho influenciado pelo intervalo de variação dos valores numéricos.
- Por isso, a transformação dos dados em uma base se torna necessária e pode ser feita de diversas formas.

Conversão Simbólico-Numérico

- Redes neurais artificiais, support vector machines e vários algoritmos de agrupamento conseguem tratar apenas dados numéricos.
- Quando aplicados sobre conjunto de dados com apresenta atributos simbólicos, os valores de tais atributos precisam ser convertidos para valores numéricos.

Conversão Simbólico-Numérico

- **Atributo do tipo nominal com apenas dois valores**: uma transformação para uma representação 0 ou 1 (binária) é suficiente.
    - Se denotam a presença ou ausência de uma característica, o valor 0 indica a ausência e o valor 1, a presença.
    - Se apresentam uma relação de ordem, o menor valor ordinal assume o valor 0 e o outro assume o valor 1.

Conversão Simbólico-Numérico

- **Atributo simbólico com mais de dois valores**: a técnica utilizada depende do atributo ser nominal ou ordinal.
- **Se não há relação de ordem entre os valores do atributo**, isso deve ser preservado para os valores numéricos gerados.
    - Isso significa que a diferença entre quaisquer dois valores numéricos deve ser a mesma.
- **Sequência de c bits**: codifica cada valor nominal como uma sequência de c bits, onde c é o número de possíveis valores ou categorias.

Conversão Simbólico-Numérico

- A codificação 1 – de – c é denominada canônica ou topológica, onde cada sequência possui apenas um bit com o valor 1 e os demais com o valor 0.
- A diferença entre dois valores pode utilizar a distância de Hamming que indica o número de posições em que as sequências apresentam valores diferentes.

Azul	100000
Amarelo	010000
Verde	001000

Conversão Simbólico-Numérico

- A distância de Hamming entre qualquer par de valores é igual a 2, ou seja, apenas duas posições do string binário têm valores diferentes.

Atributo nominal	**Código** 1–*de*–*c*
Azul	100000
Amarelo	010000
Verde	001000
Preto	000100
Marrom	000010
Branco	000001

Conversão Simbólico-Numérico

- **Se há uma relação de ordem**, o atributo é do tipo ordinal, e a codificação deve preservar essa relação.
- Nesse caso, basta ordenar os valores categóricos ordinais e codificar cada valor de acordo com sua posição na ordem.
- A distância entre os valores varia de acordo com a proximidade deles

Valor ordinal	Valor inteiro
Primeiro	0
Segundo	1
Terceiro	2
Quarto	3
Quinto	4
Sexto	5

Conversão Simbólico-Numérico

- Pode-se converter valores ordinais em valores binários através do **código cinza** ou do **código termômetro**.
- O **código cinza** é utilizado para correções de erro em comunicações digitais.
- O **código termômetro** aumenta os valores de forma semelhante ao aumento de temperatura em um termômetro analógico.

Valor ordinal	Código cinza	Código termômetro
Primeiro	000	00000
Segundo	001	00001
Terceiro	011	00011
Quarto	010	00111
Quinto	110	01111
Sexto	100	11111

Conversão Numérico-Simbólico

- Se o atributo quantitativo for do tipo **discreto e binário**, deve-se apenas associar um **nome** a cada valor.
- Se o atributo original for uma sequências binárias sem uma relação de ordem entre si, deve-se substituir por um nome ou categoria.
- Resumindo, decodificar a codificação proposta anteriormente para transformação Simbólico-Numérico considerando valores discretos e sem relação de ordem.
- Métodos de discretização são aplicados nos demais casos.

Conversão Numérico-Simbólico

- **Métodos de discretização**: transformar atributos quantitativos em qualitativos, designando valores numéricos a intervalos ou categorias.
- O conjunto de possíveis valores numéricos é separado em intervalos, convertendo cada intervalo de valores numéricos em um valor qualitativo.
- **Métodos de discretização paramétricos**: o **usuário** influência a definição dos intervalos através da **escolha de parâmetros** como número máximo de intervalos.
- **Métodos de discretização não paramétricos**: os intervalos são estabelecidos através das **informações presentes nos valores do atributo**.

Conversão Numérico-Simbólico

- **Métodos de discretização supervisionados**: utilizam a informação sobre a classe dos exemplos.
    - Costumam gerar melhores resultados, pois estabelecer intervalos sem conhecimento das classes pode levar à sua mistura.
    - Exemplo: escolher pontos de corte que maximizam a pureza dos intervalos, utilizando uma métrica como entropia.

Conversão Numérico-Simbólico

- Estratégias para transformar valores dos atributos quantitativos em valores qualitativos:
  - **Larguras iguais**: divisão do intervalo original dos valores em subintervalos com o mesmo tamanho, mas estratégia pode ser afetada por outliers.
  - **Frequências iguais**: cada subintervalo é criado com o mesmo número de objetos, podendo gerar intervalos de tamanhos bastante diferentes.
  - **Aplicação de algoritmo de agrupamento de dados**.
  - **Inspeção visual**.

Conversão Numérico-Numérico

- Aplica-se para dados com considerável **variação de valores**, ou seja, elevada distância entre os limites inferior e superior.
- Aplica-se também quando há vários atributos em **escalas diferentes** de valores.
- Exemplo 1: Dados de um atributo com valores inteiros relativos, quando apenas os valores absolutos são importantes. Converte-se todos os valores relativos para absolutos.
- Exemplo 2: Normalização de dados cujos limites de valores de atributos distintos são muito diferentes, evitando que um atributo predomine sobre outro.

Conversão Numérico-Numérico

- Nem sempre a normalização é recomendável!!
- Quando necessária, aplica-se a normalização individualmente a cada atributo
- Normalização por **amplitude** considerando:
    - **Reescala**: define-se uma nova escala de valores, limites mínimo e máximo, para todos os atributos.
    - **Padronização**: define-se um valor central e um valor de espalhamento comuns para todos os atributos.

Conversão Numérico-Numérico

- A **normalização por reescala** é conhecida como **normalização min-max:**

$$x_{ij}^{novo} = min_j^{novo} + \frac{x_{ij} - min_i(x_{ij})}{max_i(x_{ij}) - min_i(x_{ij})} (max_j^{novo} - min_j^{novo})$$

min_j^{novo} : novo valor mínimo definido para o atributo j

max_j^{novo} : novo valor máximo definido para o atributo j

$min_i(x_{ij})$: valor mínimo atual do atributo j

$max_i(x_{ij})$: valor máximo atual do atributo j

Conversão Numérico-Numérico

- A **normalização por padronização** adiciona ou subtrai uma medida de localização como a média que é dividida ou multiplicada por uma medida de escala como a variância.

$$x_{ij}^{novo} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

- Atributos diferentes terão limites inferiores e superiores distintos com os mesmos valores para as medidas de escala e espalhamento.
- A padronização **lida melhor com outliers** ao fazer com que os atributos mais importantes possuam limites maiores.

Conversão Numérico-Numérico

- A **normalização por distribuição** muda a escala de valores de um atributo.
- Exemplo: Seja X={1,5,9, 3} um conjunto de atributos.
  - X={1,5,9, 3} $\Longrightarrow$ (1,3,5,9)=(1o, 2o, 3o, 40) ordenação para **ranqueamento**
  - Normalização de X via ranqueamento: X={1,5,9, 3} $\Longrightarrow$ X^{Novo}={1,3,4,2}
  - Substituição de cada valor pela **posição** que ele ocupa no ranking!!
  - Se todos os valores originais forem distintos, o resultado é uma distribuição uniforme.

Conversão Numérico-Numérico

- **Tradução**: o valor de um atributo de um dado tipo é traduzido para um valor do mesmo tipo, mais facilmente manipulável.
    - Conversão de um atributo com data de nascimento para idade.
    - Conversão de graus Celsius para Fahrenheit.
    - Conversão da localização por um aparelho de GPS para código postal.

Redução de Dimensionalidade

- Bases de dados para aplicação de métodos de aprendizado de máquina podem apresentar um número elevado de atributos.
- Exemplo 1: Aplicações em reconhecimento de imagens podem considerar cada pixel como um atributo, logo, se cada instância de uma imagem tiver 1024 por 1024 pixels, mais de um milhão de atributos serão considerados.
- Exemplo 2: Dados de expressão gênica apresentam dezenas de amostras tratadas como objetos, enquanto os genes são os atributos. Logo, cada objeto pode conter milhares de atributos.

Redução de Dimensionalidade

- **Maldição da dimensionalidade**: o aumento na dimensão dos dados, ou seja, no número de atributos gera um aumento exponencial na demanda por dados para treinar modelos de aprendizado de máquina gerais e precisos o suficiente.
- Duas abordagens para redução de dimensionalidade:
  - Agregação;
  - Seleção de Atributos

Redução de Dimensionalidade

- **Agregação**: substitui os atributos originais por novos atributos formados pela combinação de grupos de atributos, através de funções lineares ou não lineares.
    - Análise de Componentes Principais (Principal Component Analysis - PCA) é um exemplo de método que avalia a correlação dos atributos, reduzindo a dimensionalidade do conjunto de dados original pela eliminação de redundâncias.
    - **Desvantagem**: a combinação de atributos em técnicas de agregação pode levar à perda dos valores originais. A preservação de valores pode ser relevante em áreas como biologia, finanças, medicina e monitoramento ambiental.

Seleção de Atributos

- Mantém uma parte dos atributos originais e descarta os demais atributos.
- As técnicas de seleção de atributos buscam em geral um **subconjunto ótimo** de atributos de acordo com um dado critério como as abordagens:
  - Embutida;
  - Baseada em filtro;
  - Baseada em wrapper.

Seleção de Atributos

- **Embutida:** a seleção do subconjunto é embutida ou integrada no próprio algoritmo de aprendizado.
    - Árvores de decisão realizam esse tipo de seleção interna de atributos.
- **Baseada em filtro:** há primeiro uma etapa de pré-processamento, onde um subconjunto dos atributos originais é filtrado de acordo com algum critério.
    - As heurísticas aplicadas na filtragem apresentam baixo custo computacional, tratando eficientemente um grande volume de dados.
    - Um filtro poderia considerar a correlação entre os atributos, e selecionar apenas um dos atributos entre vários altamente correlacionados.

Seleção de Atributos

- A redução baseada em filtro não leva em consideração o algoritmo de aprendizado que utilizará esse subconjunto.
- **Vantagem**: Para aplicações considerando várias técnicas de aprendizado de máquina, a independência dos filtros em relação à técnica se torna vantajosa.
- **Desvantagem**: A não interação entre filtro e técnica pode pode comprometer o desempenho como um todo.

Seleção de Atributos

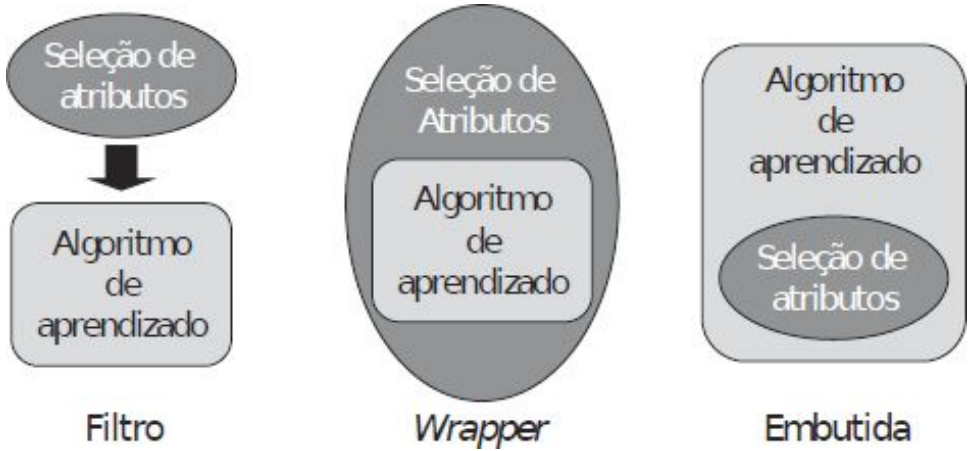
- **Baseada em wrapper:** utiliza alguma técnica de aprendizado como caixa-preta na seleção.
- A **técnica de aprendizado** pode estar associada a uma **técnica de amostragem** onde, para cada possível subconjunto de atributos, a técnica é consultada e o subconjunto que apresentar a melhor combinação entre redução da taxa de erro e redução do número de atributos é selecionado.

Seleção de Atributos

- **Vantagens**:
  - técnicas baseadas em wrapper representam uma alternativa simples e poderosa para selecionar atributos;
  - por incorporar o viés do classificador, essas técnicas conseguem obter um conjunto de atributos que melhoram o desempenho do modelo.
- **Desvantagem**: são criticadas por serem técnicas de força bruta, com um custo computacional elevado. Porém, estratégias de busca eficientes têm sido utilizadas por algumas dessas técnicas.

Seleção de Atributos

- As técnicas embutidas costumam **utilizar melhor os dados disponíveis** do que as técnicas baseadas em wrapper.
- As técnicas embutidas também são **mais rápida** já que não precisar retreinar um algoritmo de aprendizado de máquina para cada novo conjunto de atributos.
- Todavia, nem toda técnica de aprendizado realiza uma seleção de atributos embutida diretamente.



Ordenação de Atributos

- Forma simples de seleção dos atributos que **ordena** atributos baseado em sua **relevância** para determinado critério.
    - Exemplo: A classificação dos objetos nas diferentes classes é geralmente realizada de forma **univariada**, ou seja, cada atributo é avaliado independentemente dos demais. Assim, os atributos no topo da ordenação são selecionados para utilização pelo classificador.

Ordenação de Atributos

- Técnicas de ordenação ou ranking: aplicam algum critério na medida de importância dos atributos como similaridade (medidas de correlação) e diferença (medidas de distância) entre vetores.
- Tais medidas podem ser paramétricas e não paramétricas.
- **Medidas paramétricas** fazem suposições sobre a distribuição estatística das medidas dentro de cada grupo ou classe.

Ordenação de Atributos

- As **medidas não paramétricas** não fazem essa suposição e são mais robustas ao especificar uma hipótese em termos de distribuições populacionais, ao invés de parâmetros como média e desvio padrão.
- As medidas não paramétricas detectam diferenças entre populações quase tão bem quanto as medidas paramétricas quando suposições como normalidade precisam ser satisfeitas.
- Quando essas suposições não são satisfeitas, medidas não paramétricas podem ser, e frequentemente são, mais poderosas que as paramétricas para detectar diferenças entre populações.

Ordenação de Atributos

- Para o caso da **ordenação dependente de classe**, o primeiro atributo é aquele que melhor discrimina os objetos das diferentes classes, o segundo é o segundo melhor atributo para essa discriminação e assim por diante.
- Na **seleção do subconjunto**, os atributos que fazem parte do subconjunto selecionado não necessariamente estariam no topo da lista se uma técnica de ordenação fosse utilizada.

Ordenação de Atributos

- A forma como os atributos selecionados atuam de forma coletiva, em conjunto, impacta mais na seleção do subconjunto.
- Isso ocorre, por exemplo, porque dois atributos situados próximos na lista ordenada podem estar correlacionados.
- Logo, a ordenação não é capaz de detectar redundâncias entre os atributos.

Seleção de Subconjuntos

- Trata-se de um processo **computacionalmente mais custoso** que a ordenação dos atributos à medida que cresce o número de atributos.
- Uma alternativa é primeiro ordenar os atributos originais, aplicando uma técnica de ordenação, selecionando em seguida um subconjunto a partir daqueles melhor classificados pela ordenação.

Seleção de Subconjuntos

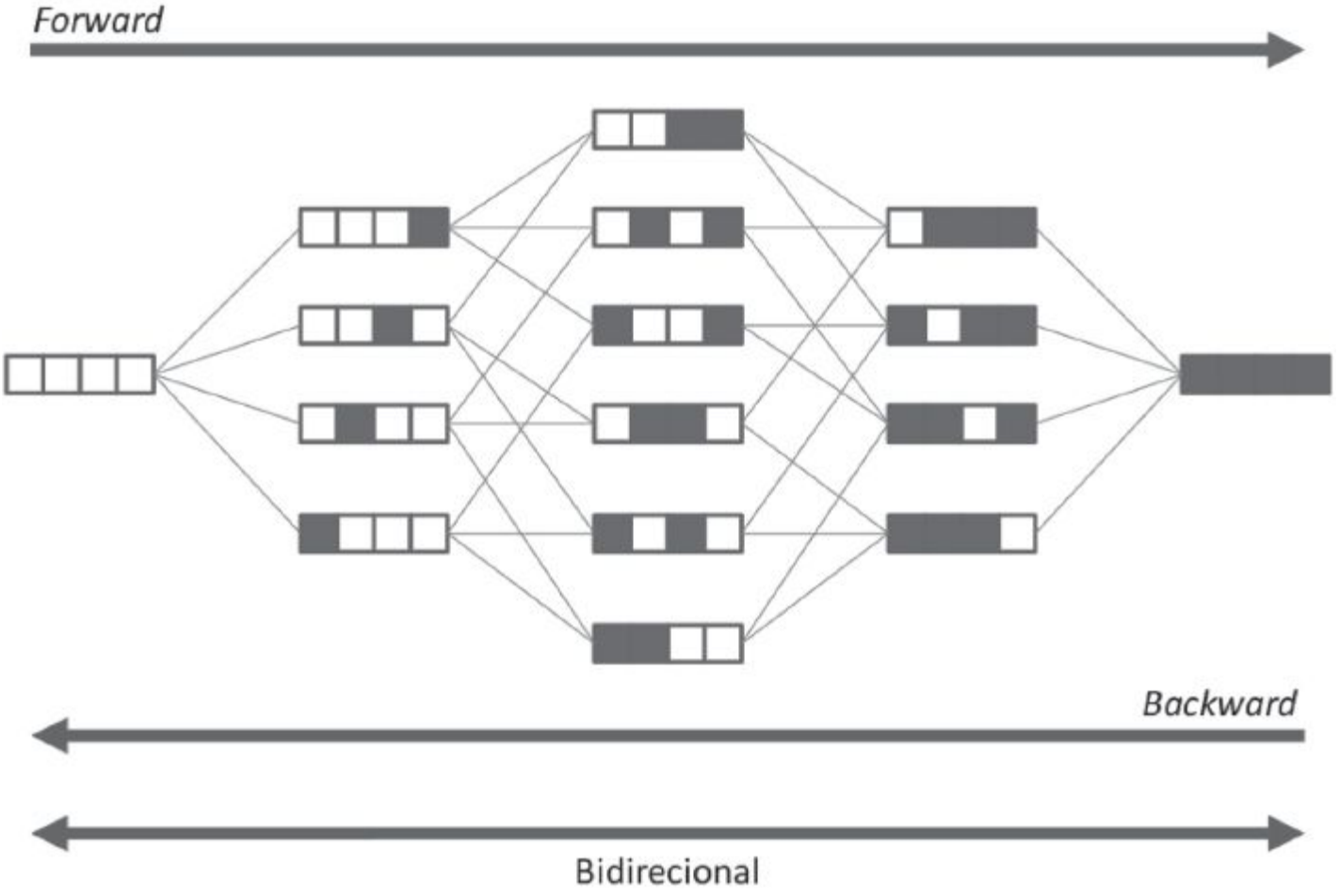
- A seleção de um subconjunto de atributos pode ser tratada como um problema de busca, onde cada ponto no espaço de busca seria um possível subconjunto de atributos.
- Nesse espaço de busca, os **critérios de avaliação** dos pontos (subconjunto de atributos) seriam técnicas apresentadas como **filtro**, **wrapper** ou **embutida**.

Seleção de Subconjuntos

- O **ponto de partida e direção da busca** pode ser executado considerando:
  - **Geração para trás (backward generation)**: seleciona todos os atributos e remove um por vez.
  - **Geração para a frente (forward generation)**: inicia sem atributo e inclui um atributo por vez.

Seleção de Subconjuntos

- O **ponto de partida e direção da busca** pode ser executado considerando:
  - **Geração bidirecional (bidirectional generation)**: inicia em qualquer ponto, adicionando e removendo atributos.
  - **Geração aleatória (random generation)**: o ponto de partida da busca e atributos a serem removidos ou adicionados são decididos aleatoriamente.


Forward
Backward
Bidirecional

Seleção de Subconjuntos

- Considerando a **estratégia de busca** a ser executada:
  - **Busca completa (exponencial ou exaustiva):** avalia todos os possíveis subconjuntos.
  - **Busca heurística (sequencial):** aplica regras e métodos para conduzir a busca, não garantindo que uma solução ótima seja encontrada.
  - **Busca não determinística:** aplica-se uma geração estocástica na busca, também sem garantias de que uma boa solução ou a melhor solução possível possa ser encontrada antes do final da busca.

Seleção de Subconjuntos

- Considerando o **critério de parada** da busca:
  - **Busca exaustiva**: a busca é encerrada quando todos os subconjuntos forem testados.
  - **Melhor subconjunto encontrado**: termina a busca pelo melhor subconjunto de atributos a partir de um número máximo de alternativas testadas, um número de atributos a serem selecionados sem degradação do desempenho do classificador ou o tempo de processamento.