DEPARTAMENTO DE COMPUTAÇÃO
E TECNOLOGIA

Sistemas de Informação
UFRN

Pré-processamento dos Dados

Karliane Vale (karliane.vale@ufrn.br)
Huliane Medeiros (hulianeufrn@gmail.com)

Contextualização

- Apesar do algoritmos serem usados constantemente para extrair conhecimento dos dados, seu desempenho é geralmente afetado pelo estado dos dados.

- Por exemplo: diferentes características, dimensões, formatos, atributos numéricos ou simbólicos, dados estar limpos ou conter ruídos e imperfeições, com valores incorretos, inconsistentes, duplicados ou ausentes, poucos ou muitos objetivos, assim como ter um número pequeno ou elevado de atributos.

Objetivo das técnicas de pré-processamento de dados

- Usadas para melhorar a qualidade dos dados, por meio da eliminação ou minimização dos problemas

- Tornar os dados mais adequados para sua utilização por um algoritmo de AM
  - Por exemplo: alguns algoritmos de AM trabalham apenas com valores numéricos

Vantagens das técnicas de pré-processamento de dados

- Pode facilitar o uso dos algoritmos de AM, levar a construção de modelos mais fiéis à distribuição real dos dados , reduzindo a sua complexidade

- Tornar mais fáceis e rápidos o ajuste de parâmetro do modelo e seu posterior uso

- Pode facilitar a interpretação dos padrões extraídos pelo modelo

Eliminação manual de atributos

Conjunto de atributos que farão parte da análise é geralmente definido de acordo com o experiência de especialistas do domínio dos dados

Especialistas pode decidir que alguns atributos são irrelevantes para o diagnóstico clínico

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

Eliminação manual de atributos

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

Eliminação manual de atributos

- Existem outras situações em que um atributo irrelevante pode ser facilmente identificado
  - Por exemplo: um atributo possui o mesmo valor para todos os objetos. Não contém informação que ajude a distinguir os objetos

- Importante: um atributo não precisa ter o mesmo valor para todos os objetos para ser considerado irrelevante
  - Técnicas de seleção de atributos podem ser utilizadas para eliminar atributos irrelevantes

Integração de Dados

- Podem estar distribuídos em diferentes conjuntos de dados
- Necessário integrar
- Na integração, é necessário identificar quais são os objetos que estão presentes nos diferentes conjuntos de dados a serem combinados
- Conhecido como problema de identificação de identidade

Integração de Dados

- Essa identificação é realizada por meio da busca por atributos comuns nos conjuntos a serem identificados
- Por exemplo: conjunto de dados médicos podem ter um atributo que identifica o paciente
  - os objetos de diferentes conjuntos que possuem o mesmo valor para o atributo que identifica o paciente são combinados em um único objeto do conjunto integrado

Integração de Dados

- Aspectos que podem dificultar a integração:
  - atributos correspondentes podem ter nomes diferentes em diferentes bases de dados
  - dados atualizados em momentos diferentes
- Metadados são usados para minimizar esses problemas
  - São dados sobre os dados, que as descrever as principais características, podem ser utilizados para evitar erros no processo de integração

Amostragem de dados

- Algoritmos de AM podem ter dificuldade em lidar com um grande número de objetos

- Associado ao número de objetos, existem um balanço entre eficiência computacional e acurácia.

- Quando mais dados são utilizados, maior tende a ser a acurácia do modelo e menor a eficiência computacional do processo indutivo, pois um número muito grande de objetos pode tornar o tempo de processamento muito longo.

Amostragem de dados

- Para se obter um bom compromisso entre eficiência e acurácia, trabalha-se com uma amostra ou subconjuntos dos dados
- Umas amostra leva ao mesmo desempenho obtido com o conjunto completo, porém com um custo computacional muito menor.
- Uma amostra pequena pode não representar bem o problema que se deseja modelar
- A amostra deve ser representativa do conjunto de dados original

Amostragem de dados

- Se for muito grande, são reduzidas as vantagens de utilizar amostragens

- O ideal é que a amostra não seja grande, mas que seus dados obedeçam à mesma distribuição estatística que gerou o conjunto de dados original

Amostragem de dados

- Amostragem aleatória simples:
  - amostragem simples sem reposição de exemplos, quando cada exemplo pode ser selecionado apenas uma vez
  - amostragem com reposição, quando uma cópia dos exemplos selecionados é mantida no conjunto de dados original

Amostragem de dados

- Amostragem estratificada: usada quando as classes apresentam propriedades diferentes, por exemplo, números de objetos bastante diferentes
  - manter o mesmo número de objetos em casa classe
  - manter o mesmo número de objetos em casa classe proporcional ao número de objetos da classe no conjunto original

Amostragem de dados

- Amostragem progressiva: começa com uma amostras pequenas e aumenta progressivamente o tamanho da amostra extraída, enquanto a acurácia preditiva continuar a melhorar

Dados desbalanceados

- Redefinir o tamanho do conjunto de dados
- Utilizar diferentes custos de classificação para as diferentes classes
- Induzir um modelo para uma classe

Limpeza de dados

- 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

Limpeza de dados

- Dados inconsistentes: possuem valores conflitantes em seus atributos

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

- Dados redundantes: quando um objeto é muito semelhantes a um outro objeto do mesmo conjuntos de dados

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 com ruídos

- Dados que aparentemente não pertencem a distribuição dos dados

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