Mineração de Dados em Biologia Molecular

Pré-processamento de dados

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

Tópicos

- Introdução
- Qualidade de dados
  - Fontes de problemas
- Limpeza de dados
- Desbalanceamento
- Transformação de dados

Pré-processamento

- Prepara os dados para seu uso por algoritmos de AM
- Procura melhorar desempenho do algoritmo
  - Custo
    - Tempo
    - Memória
  - Qualidade da modelo gerado
    - Acurácia preditiva

Exemplo

- Primeiro passo:
  - Eliminar atributos irrelevantes

Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
Pedro	não	não	baixo	não	1100	saudável
Maria	sim	sim	alto	não	600	saudável
José	sim	não	baixo	sim	2000	doente
Ana	sim	não	alto	sim	1800	saudável
Leila	não	não	alto	sim	900	doente

Qualidade de dados

- Em geral, dados não foram gerados para uso em AM
  - Produzidos para outros propósitos
  - Frequentemente apresentam problemas
- Algoritmos de AM precisam geralmente de dados "limpos"
  - Entra lixo, sai lixo
  - Problemas nos dados precisam ser detectados e corrigidos
    - Limpeza de dados

Qualidade de dados

- Problemas nos dados podem ter causa:
  - Sistemática (determinística)
    - Mais fácil de detectar e corrigir
  - "Aleatória"

Possíveis causas de problemas

- Falha humana
- Má fé
- Falha no processo ou dispositivo de coleta ou de medição de dados
- Limitações do dispositivo de coleta ou de medição
- Mudança de conceito

Possíveis consequências

- Valores de atributos preditivos podem ser perdidos ou modificados
- Obtenção de objetos que sejam:
  - Espúrios ou duplicados
    - Ex.: diferentes registros para mesma pessoa que morou em endereços diferentes
  - Inconsistentes
    - Ex.: engenheiro com 3 anos de idade

Limpeza

- 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*

Valores ausentes

- Dados faltosos, faltantes, incompletos
- Várias técnicas de AM não foram projetadas com capacidade para lidar com valores ausentes

Mecanismos de ausência

- Ausência completamente aleatória (MCAR)
  - Probabilidade da ausência de valor de uma variável independe dela e de qualquer outra variável
- Ausência aleatória (MAR)
  - Probabilidade da ausência de valor de uma variável independe dela, mas depende de outras variáveis
- Ausência não aleatória (MNAR)
  - Probabilidade da ausência de valor de uma variável depende apenas dela, não de outras variáveis

Valores ausentes

- Não é raro um objeto não ter valores para um ou mais atributos
- Possíveis causas:
  - Atributo não foi considerado quando os primeiros dados foram coletados
  - Desconhecimento do valor do atributo por ocasião do preenchimento
  - Distração, mal entendido ou declinação na hora do preenchimento
  - Problema com dispositivo / processo de coleta de valores para o atributo

Exemplo de valores ausentes

Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
	não	não	baixo	não	1100	saudável
Maria	sim	sim		não	600	saudável
José	sim	não	baixo	sim		doente
Sérgio	não	não	baixo	não	1100	saudável
Ana	sim	não	alto	sim	1800	saudável
Leila		não	alto		900	doente
Marta	sim	não	baixo	sim	2000	doente

Lidar com valores ausentes

- Ignorar valores ausentes
  - Utilizar apenas os valores que estão presentes
    - Ex.: Menos atributos no cálculo da distância entre objetos
  - Modificar algoritmo de modelagem para lidar com valores ausentes
- Descartar objetos com atributos sem valores
- Preencher valores ausentes

Descarte de objetos

- Geralmente empregado quando:
  - Um dos atributos ausentes é o atributo classe
  - Objeto tem muitos valores ausentes
- Não é indicado quando:
  - Ocorre com poucos atributos do objeto
  - Há risco de perder dados importantes

Preenchimento de valor

- Criação de um novo valor que significa ausência
  - Valores nominais (sem ordem)
- Criação de um novo atributo preditivo
  - Marcando objetos em que um dado atributo tinha valor ausente
- Estimativa de um valor para o valor preditivo ausente

Estimativa do valor

- Medida de localidade
  - Média (mediana, moda) dos valores do atributo
    - Todos os valores
    - Dos objetos mais próximos e/ou da mesma classe
  - Para série temporais, medida de localidade entre valores anterior e posterior

Estimativa do valor

- Induzir valor induzido por algum estimador
  - Valor presente em objetos semelhantes
  - Utilizar algoritmo de AM
  - Alternativa mais eficiente

Valores ausentes

- Observações
  - Em alguns casos, a ausência de valor é uma informação importante sobre o objeto
  - Existem situações em que o valor pode ou precisa estar ausente
    - Ex.: Atributo número do apartamento para quem mora em uma casa
    - Ao invés de ausente, é um valor inexistente
    - Difícil tratar de forma automática
      - Criação de um novo atributo

Exercício

- Tratar dos valores ausentes da tabela abaixo

Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador	Médio	70	180	3000	adimplente
Lia		Superior	200	174	7000	inadimplente
Maria	Advogado	Médio		180	600	adimplente
José	Médico	Superior	100		2000	inadimplente
Sérgio	Bancário		82	178	5000	inadimplente
Ana	Professor	Fundam.	77	188	1800	adimplente
Luísa	Médico	Superior	100	36	2000	inadimplente
José	Médico	Médio	340		800	

Valores inconsistentes

- Dados podem conter valores inconsistentes
  - Atributos preditivos
    - Ex. Código postal invalido para uma cidade
      - Erro / engano
      - Proposital (fraude)
  - Atributo alvo
    - Podem levar a objetos conflitantes (ambiguidade)
      - Ex.: valores iguais para atributos preditivos e diferentes para atributo alvo
    - Podem ser causados por erros no processo de rotulação

Valores inconsistentes

- Algumas inconsistências são de fácil detecção
  - Violação de relações conhecidas entre atributos
    - Ex.: Valor de atributo A é sempre menor que valor de atributo B
  - Valor inválido para o atributo
    - Ex.: altura com valor negativo
  - Em outros casos, informações adicionais precisam ser consideradas
- Podem indicar presença de ruído

Exemplo de objetos inconsistentes

Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
Pedro	não	não	baixo	não	1100	saudável
Maria	sim	sim	alto	não	600	saudável
José	sim	não	baixo	sim	2000	doente
Sérgio	não	não	baixo	não	1100	doente
Ana	sim	não	alto	sim	1800	saudável
Leila	não	não	alto	sim	900	doente
Marta	sim	não	alto	sim	3000	doente

Exemplo de atributos inconsistentes

Nome	Idade	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	30	sim	baixo	sim	1000	doente
Pedro	42	não	baixo	não	1100	saudável
Maria	27	sim	alto	não	600	saudável
José	4	não	baixo	sim	2000	doente
Sérgio	38	não	baixo	não	1100	doente
Ana	63	não	alto	sim	1800	saudável
Leila	22	não	alto	sim	900	doente
Marta	53	não	alto	sim	3000	doente

Objetos redundantes

- Objetos ou atributos preditivos (quase) duplicados
  - Não trazem informação nova
  - Ex.: Pessoas em diferentes BDs com mesmo nome, mas endereço com pequenas diferenças
    - Diferença real ou erro no preenchimento
- Deduplicação
  - Detectar e eliminar (ou combinar) duplicações
  - Cuidado para não eliminar ou combinar objetos ou atributos que representam dados diferentes

Exemplo

- objetos redundantes

Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
Segio	não	não	baixo	não	1100	saudável
Maria	sim	sim	alto	não	600	saudável
José	sim	não	baixo	sim	2000	doente
Sérgio	não	não	baixo	não	1100	saudável
Ana	sim	não	alto	sim	1800	saudável
Leila	não	não	alto	sim	900	doente
Marta	sim	não	baixo	sim	2000	doente

Exercício

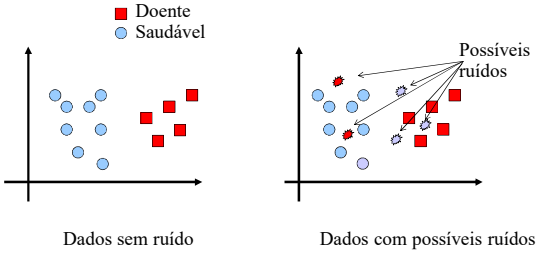
- Definir problemas existentes na tabela abaixo:

Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador		70	180	3000	adimplente
Lia	Médico	Superior	200	174	7000	inadimplente
Maria	Advogado	Médio	90	180	600	adimplente
José	Médico	Superior	200	174	7000	inadimplente
Sérgio	Bancário	Superior	82	178	5000	inadimplente
Ana	Professor	Fundam.	77	188	1800	adimplente
Luísa	Médico	Superior	100	-6	2000	inadimplente

Ruídos

- Podem levar a um superajuste do modelo obtido por um algoritmo de AM
- Difícil ter certeza que um valor é ruído
  - Tem-se apenas um indício
    - A menos que valor seja inconsistente
  - Se identificados, podem ser tratados como valores ausentes
- Nos atributos preditivos ou no atributo alvo
  - Consequências diferentes

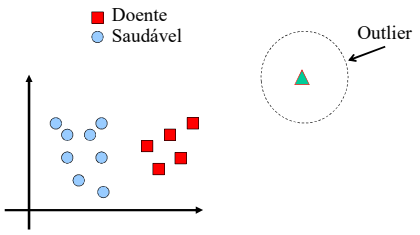
Exemplo



Outliers

- Objetos ou valores anômalos
  - Objetos que têm características diferentes da grande maioria dos demais objetos
    - Valor(es) de um ou mais atributos que destoa(m) dos valores típicos
- *Outliers* podem sugerir a presença de ruído ou ser valores legítimos
  - Em várias aplicações, objetivo é encontrar *outliers*

Outliers



Dados desbalanceados

- Número de objetos varia para as diferentes classes
  - Natural ao domínio
  - Problema com geração / coleta de dados
- Várias técnicas de AM não conseguem lidar com esse problema
  - Tendência a classificar na(s) classe(s) majoritária(s)

Dados desbalanceados

- Alternativas
  - Alteração do conjunto de dados
    - Balanceamento artificial
  - Utilizar diferentes custos de classificação para as diferentes classes
  - Induzir um modelo para uma das classes
  - Alteração ou projeto de algoritmos para lidar com desbalanceamento

Balanceamento artificial

- Redefinir o tamanho do conjunto de dados
  - Acrescentar objetos (sobreamostragem)
    - Replicar objetos da classe minoritária
    - Não adiciona informação
  - Eliminar objetos (subamostragem)
    - Ignorar objetos da classe majoritária
    - Remove informação
  - Abordagem híbrida

Dados desbalanceados

- Algumas técnicas são insensíveis ao balanceamento artificial
  - Ex.: AD e NB
  - Mas algumas vezes funciona
- Realizado para os dados de treinamento
  - Medida de avaliação de desempenho deve considerar desbalanceamento em dados de teste

Modelo para classe

- Indução de modelo para uma das classes
  - Classificação com apenas uma classe
    - Classe majoritária
    - Classe minoritária
    - Cada uma das classes

Dados desbalanceados

- Alguns problemas em algoritmos de AM só aparecem quando os dados estão desbalanceados
- Atenção!!!
  - Pode ser que uma distribuição igualitária das classes não seja boa
    - Mesmo se a população apresentar essa distribuição

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

Qualitativos para quantitativos

- Algumas técnicas trabalham apenas com valores numéricos
- Conversão depende de:
  - Existência de ordenação dos valores
    - Se existe (ordinal), mantém
    - Se não existe (nominal), não inserir
  - Número de valores
    - Se igual a 2 (binários) ou maior que 2

Conversão de valor ordinal

- Codificar para valor inteiro positivo
  - Ex. Pequeno: 1, médio: 2 e grande: 3
- Algumas técnicas trabalham apenas com valores quantitativos binários
  - Binarização

Binarização de ordinal

- Transformação no sistema numérico binário correspondente?
  - Perde ordenação
    - Valores consecutivos devem diferir em 1 bit
- Codificar cada valor por um vetor binário que mantém ordenação
  - Código cinza: 000, 001, 011, 010, ...
  - Código termômetro: 001, 011, 111

Código cinza

- Existem vários códigos cinza
  - Não é único
- Um código cinza para 3 bits:
  - 000, 001, 011, 010, ...
- Um código cinza para 2 bits:
  - 00, 01, 11, 10

Dígito	Binário	Código cinza
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111
11	1011	1110
12	1100	1010
13	1101	1011
14	1110	1001
15	1111	1000

Algoritmo código cinza

1 Começa com todos os bits iguais a zero
2 Para cada novo número
Mudar o valor do bit mais a direita que gera uma nova sequência de bits

Código termômetro

- Utiliza mais bits que código cinza
  - Tamanho cresce linearmente com número de valores

Dígito	Binário	Código termômetro
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0111
4	0100	1111

Conversão de valor nominal

- Transforma para valor quantitativo
  - Não deve inserir relação de ordem
- Codificação binária nominal sem relação de ordem
- Codificações
  - 1-de-n
  - m-de-n

Conversão de valor nominal

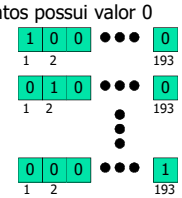
- Codificação 1-de-n
  - Codificação canônica
  - Fácil calcular moda = posição com maior número de valores 1
  - Quantidade de valores pode gerar vetores longos
- Codificação m-de-n
  - Dos n valores, m são iguais a 1 e os demais 0
  - Vários códigos

Conversão de valor nominal

- Número de valores pode ser muito grande
- Pseudo atributos
  - Cria valores novos, artificiais
- Ex.: Atributo é nome de país
  - Existem 193 países (192 representados na ONU + Vaticano)
  - Alternativa de codificação:
    - Transformar valores nominais em valores numéricos utilizando a codificação 1-de-n

Alternativa 1

- Transformar valores nominais em valores binários utilizando a codificação 1-de-n
  - Maldição da dimensionalidade
  - Grande parte dos elementos possui valor 0
    - Valores esparsos



Alternativa 2

- Transformar 193 atributos em 4 (10) pseudo-atributos
  - Continente: 7 valores binários
  - IDH: 1 valor real
  - População: 1 valor inteiro
  - Área: 1 valor inteiro

Exercício

- Transformar valores do atributo nome de automóvel em pseudo-atributos
  - Ex.: Uno, fox, amarok, corsa, zafira, corolla, TR4, gol, palio, doblo, clio, kangoo, omega

Quantitativos para qualitativos

- Discretização de valores
  - Transformar valores numéricos em intervalos (ou categorias)
- Subtarefas
  - Definição do número de categorias
    - Geralmente feito pelo usuário
  - Definição de como mapear valores dos atributos numéricos para essas categorias
    - Por frequência ou largura dos intervalos
    - Geralmente feito por um algoritmo

Transformação de atributos

- Muda valor numérico de um atributo para outro valor numérico
  - Limites de valores para atributos distintos podem ser muito diferentes
    - Evitar que um atributo predomine sobre outro
      - A menos que isso seja importante
  - Valores podem estar concentrados em uma determinada faixa ou região
  - Possível necessidade de binarização

Transformação de atributos

- Aplicada aos valores de um atributo específico para todos os exemplos
- Variações
  - Funções simples
  - Normalização
  - Padronização

Funções simples

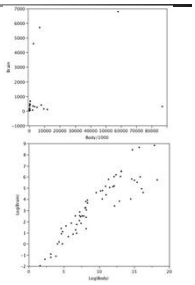
- Uma função matemática simples é aplicada a cada valor do atributo
  - Muda distribuição de valores de um atributo
  - Possíveis transformações para um atributo x de um conjunto de dados:
    - x^k, $log(x)$, e^x, $\sqrt{x}$, $1/x$, $sqrt(x)$, $seno(x)$ e $|x|$

Funções simples

- Valor absoluto
  - Em algumas aplicações, apenas magnitude do valor de um atributo é importante
  - Converte valor de todos os atributos para o valor positivo correspondente
    - Ex.: -4, 5 e –2 se tornam 4, 5 e 2

Funções simples

- Utilizando função $\log_{10}$
  - Comprime valores de atributos com um grande intervalo de possíveis valores
  - Ex.: relação, para alguns animais, entre:
    - Peso do cérebro e
    - Peso do corpo



http://onlinestatbook.com/2/transformations/log.html

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]

$$x' = \frac{(x - \min_x)}{(\max_x - \min_x)}$$

Exercício

- Normalizar os valores 12, 5, 4, 10, 20, 3 para os intervalos:
  - [-1, +1]
  - [-7, 12]

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)

$$x' = \frac{(x - \bar{x})}{\sigma}$$

Exercício

- Converter os seguintes valores numéricos utilizando normalização e padronização

Valores	Re-escala	Padronização
3		
9		
5		
11		
5		
7		

$$\text{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Exercício

- Converter os dados abaixo para valores numéricos no intervalo [0, 1]

Febre	Enjoo	Batimentos	Vacina	Diagnóstico
baixa	sim	baixo	A	doente
média	não	normal	C	saudável
alta	sim	alto	B	saudável
alta	não	baixo	A	doente
baixa	não	alto	D	saudável
média	não	sem	C	doente

Conversão de valores numéricos

- É preferível padronizar a normalizar
- Em algumas aplicações
  - Atributos mais importantes podem ser deixados com limites maiores

Tradução

- 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

Considerações finais

- Qualidade de dados
  - Fontes de problemas
- Pré-processamento
- Limpeza de dados
- Desbalanceamento
- Transformação de dados

Perguntas