DEPARTAMENTO DE COMPUTAÇÃO
E TECNOLOGIA

Sistemas de Informação
UFRN

Pré-processamento dos Dados

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

Técnicas de pré-processamento

- Eliminação manual de atributos
- Integração de dados
- Amostragem de dados
- Dados desbalanceados
- Limpeza de dados
- Transformação de dados
- Redução de dimensionalidade

Dados desbalanceados

- Por exemplo: conjunto de dados de clientes de um banco
  - cada cliente é rotulado como tendo ou não ficado com saldo negativo nos últimos 90 dias
  - porcentagem de clientes com saldo negativo = 5%
  - classe majoritária terá 95% dos dados

- Vários algoritmos têm o desempenho prejudicado

- Tende a favorecer a classificação de novos dados na classe majoritária

Dados desbalanceados

- Redefinir o tamanho do conjunto de dados:
  - acréscimo de objetos na classe minoritária ou eliminação na classe majoritária

- Problemas com o acréscimo:
  - risco de objetos acrescentados representarem situação que nunca ocorrerão, induzindo ao erro
  - *overfitting*, modelo superajustado aos dados de treinamento

Dados desbalanceados

- Problemas com a eliminação:
  - possível que dados importantes para a indução do modelo correto sejam perdidos
  - *underfitting*, modelo induzido não se ajusta aos dados de treinamento

Dados desbalanceados

- Utilizar diferentes custos de classificação para as diferentes classes:
  - dificuldade em definir os custos. Número de objetos da classe majoritária for o dobro de exemplos da minoritária, um erro de classificação para um exemplo da classe minoritária pode equivaler à dois erros de classificação para um exemplo da classe majoritária
  - definição dos diferentes custos não é trivial. Dificuldade de incorporar a consideração de diferentes custos em alguns algoritmos
  - baixo desempenho quando os elementos da classe majoritária são semelhantes

Dados desbalanceados

- Induzir um modelo para uma classe:
  - As classes minoritárias ou majoritárias são aprendidas separadamente

Transformação de dados

- Técnicas limitadas à manipulação de valores de determinados tipo, por exemplo, apenas numéricos, apenas simbólicos

  - Conversão simbólico-numérico
  - conversão numérico-simbólico
  - transformação de atributo numéricos

Transformação de dados

- Conversão simbólico-numérico
  - redes neurais e SVM lidam apenas com valores numéricos
  - atributos nominal e assume apenas dois valores:
    - 0 ausência e 1 presença
    - relação de ordem, 0 menor valor ordinal, 1 maior valor ordinal
  - atributos simbólicos com mais de dois valores:
    - se não tiver relação de ordem, a inexistência dessa relação deve prevalecer

Transformação de dados

- Conversão simbólico-numérico: codificação 1 - de - *c*

Codificar cada valor nominal por uma sequência de *c* bits. *c* é igual o número de possíveis valores ou categorias

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

Dependendo do número de valores, a sequência pode ficar muito longa. Codificar os nomes dos países

Transformação de dados

- Usar pseudoatributo

Pseudoatributo	#Valores
Continente	7 (b)
PIB	1 (i)
População	1 (i)
TMA	1 (i)
Área	1 (i)

Valores do tipo binário, inteiro ou real

Transformação de dados

- Existe relação de ordem, atributo do tipo ordinal e a codificação deve preservar essa relação

- Deve usar uma codificação que a ordem esteja clara

- Valor numérico é um número inteiro ou real, transformação simples e direta

Transformação de dados

- Ordenar os valores categóricos ordinais e codificar cada valor de acordo com sua posição na ordem

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

Transformação de dados

- Se for necessário transformar valor ordinal em binário

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

Transformação de dados

- Conversão numérico-simbólico
  - alguns algoritmos trabalham melhor com valores qualitativos
  - atributo quantitativo do tipo discreto e binário, com apenas dois valores, basta associar o nome a cada valor
  - atributo formado por sequência binária, sem relação de ordem, cada sequência pode ser substituído por um nome ou categoria
  - demais casos, usa técnicas de discretização, transforma valores numéricos em intervalos ou categorias

Transformação de dados

- Transformação de atributos numéricos
  - limite inferior e superior dos valores são muito diferentes
  - muitos atributos em escalas diferentes
  - transforma para evitar que um atributo predomine o outro
  - normalização
    - amplitude
    - distribuição

Transformação de dados

- Transformação de atributos numéricos
  - normalização por amplitude
    - reescala: define uma nova escala de valores, limites mínimos e máximos para todos os atributos
    - padronização: define um valor central e um valor de espalhamento comuns para todos os atributos

Transformação de dados

- Transformação de atributos numéricos
  - normalização por distribuição
    - muda a escala de valores dos atributos
    - aplicação da função ordenar os valores do atributo a ser normalizado e a substituição de cada valor pela posição que ocupa o *ranking*
      - VALORES: 1, 5, 9 e 3, respectivamente, 1, 3, 4 e 2

Redução da dimensionalidade

- Exemplo: aplicações de reconhecimento de imagens
  - Imagem com 1024 por 1024 pixels - mais de um milhão de atributos

- Maldição da dimensionalidade
  - Efeitos negativos causados pelo aumento arbitrário do número de atributos na classificação

Maldição da dimensionalidade

- A quantidade de dados de que você precisa, para alcançar o conhecimento desejado, impacta exponencialmente o número de atributos necessários

- Devemos entender o termo **Maldição da Dimensionalidade** se referindo a vários fenômenos que surgem na análise de dados em espaços com muitas dimensões (atributos), muitas vezes com centenas ou milhares de dimensões. Lembrando que, basicamente, adicionar características não significa que sempre melhora o desempenho de um classificador

- Na prática: implica que para um dados tamanho de amostras, existe um número máximo de características a partir do qual o desempenho do classificador irá degradar, ao invés de melhorar

Maldição da dimensionalidade

- Exemplo: *dataset* com 1.000 atributos por 200.000 registros (instâncias), humanamente é impossível analisar todas essas informações e para uma algoritmo também é muito custoso

- Digamos que dos 1.000 atributos, apenas 100 são relevantes, ou seja os demais são atributos ruins ou correlacionados.
  - Se aplicamos o *dataset* com todos os atributos, por exemplo, em um algoritmo K-NN, o resultado será um mau desempenho na classificação, pois o algoritmo K-NN normalmente é enganado quando o número de atributos é grande.
  - Outros classificadores também tem o seu desempenho prejudicado.

Redução do número de atributos

- Melhora o desempenho do modelo induzido
- Reduz o custo computacional
- Torna os resultados obtidos mais compreensíveis
- Técnicas para redução do número de atributos
  - Agregação
  - Seleção de atributos

Técnicas para redução do número de atributos - AGREGAÇÃO

- Substituem os atributos originais por novos atributos formados pela combinação de grupos de atributos (perdem os valores originais)
- Combinam os atributos originais por meio de funções lineares ou não
  - Análise de Componentes Principais (PCA)
    - descorrelaciona estatisticamente os exemplos, reduzindo a dimensionalidade do conjunto de dados origianl pela eliminação de redundâncias

Técnicas para redução do número de atributos - SELEÇÃO DE ATRIBUTOS

- Mantêm uma parte dos atributos originais e descartam os demais atributos
- Na prática, vários atributos passíveis de eliminação não são facilmente identificados
  - Por isso o uso de técnicas visuais se torna pouco eficiente
    - Número muito grande de instâncias e/ou atributos
    - Relações complexas entre atributos, cuja descoberta é difícil
- Técnicas automáticas: procuram por um subconjunto ótimo de atributos de acordo com um dado critério

Abordagens para seleção automática de atributos

Filtro

Wrapper

Embutida

Filtro

- Um subconjunto dos atributos originais é filtrado de acordo com algum critério
  - Não leva em consideração o algoritmo de AM que utilizará esse subconjunto
- A independência dos filtros em relação ao algoritmo de AM pode ser vantajosa
  - Se houver a necessidade de os atributos selecionados serem empregados com diversos algoritmos de AM
- Heurísticas utilizadas para filtragem são de baixo custo computacional
  - Podem ser rápidos
  - Capazes de lidar eficientemente com uma grande quantidade de dados

Wrapper

- Utiliza algum algoritmo de AM como uma caixa preta para a seleção
  - Geralmente junto com uma técnica de amostragem

- Para cada possível subconjunto de atributos o algoritmo é consultado
  - O subconjunto que apresentar a melhor combinação entre redução da taxa de erros e redução do número de atributos é selecionado

- Técnica simples e poderosa, mas com alto custo computacional

- Em geral, conseguem obter um conjunto de atributos que leva a um melhor desempenho posterior do modelo

Embutida

- A seleção do subconjunto de atributos é embutida ou integrada no próprio algoritmo de AM
  - Exemplo: árvore de decisão

- As técnicas embutidas fazem melhor uso dos dados disponíveis do que as baseadas em wrapper

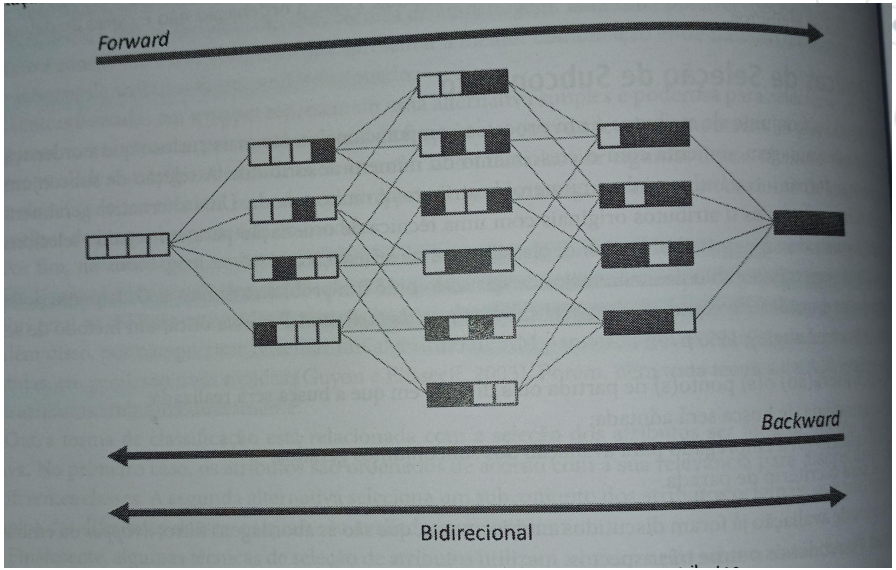
- Em geral são mais rápidas por não precisar retreinar um algoritmo de AM para cada novo conjunto de atributos

Técnicas de ordenação (ou *ranking*)

- Pode ser vista como uma forma simples de ordenação
  - Os atributos são ordenados de acordo com sua relevância para um dado critério
  - A ordenação é realizada de maneira univariada - cada atributo é avaliado independentemente dos demais
- Usam critérios como similaridade (correlação) ou diferença (distância) para medir a importância dos atributos
  - Medidas paramétricas fazem suposições sobre a distribuição estatística das medidas dentro de cada grupo ou classe (ex.: média e desvio padrão)
  - Medidas não paramétricas não fazem suposições e são mais robustas (ex.: especificam uma hipótese em termos de distribuições populacionais

Técnicas de seleção de subconjuntos

- Pode ser vista como um problema de busca
  - cada ponto no espaço de busca pode ser considerado como um possível subconjunto de atributos



Material complementar

- Canal no youtube - Carlos Alexandre Fernandes https://www.youtube.com/channel/UCrxDYm1T9Y1uXc_0X7sOekA/videos