CIÊNCIA DE DADOS - 05

Prof. Júlio Cesar Nievola

PPGIa – PUCPR

08/junho/2019

Normalização – 1

- Alguns métodos, como os baseados no cálculo da distância entre pontos, podem precisar de normalização para obter os melhores resultados
- Os valores medidos são escalados para uma faixa específica
- *Escalamento decimal*: move o ponto decimal mas preserva os dígitos:

$$v'(i) = v(i)/10^k$$

Normalização – 2

- *Normalização mín-máx*:

$$v'(i) = \frac{v(i) - min[v(i)]}{max[v(i)] - min[v(i)]}$$

min e *max* para a característica *v* são calculados automaticamente do conjunto ou dados pelo especialista

- *Normalização pelo Desvio Padrão*:

$$v'(i) = \frac{v(i) - mean(v)}{std(v)}$$

mean(v) e *sd*(v) são calculados do conjunto de dados

Suavização de Dados

- Pequenas diferenças no valor das instâncias de uma variável podem não ser significativas e degradar o desempenho e o resultado final
  - Suavização simples: arredondamento de valores
  - Reduzir o número de valores distintos reduz a dimensionalidade
  - Discretizar variáveis contínuas

Diferenças e razões

- A faixa de valores para variação relativa em geral é menor que para valores absolutos (usar diferença – chamado de *gradiente*)
- O nível de aumento/diminuição de uma variável (razão) pode melhorar o desempenho (e.g. dados médicos, uso de body-mass index - BMI em lugar de peso e altura)
- Relações lógicas (A > B) podem ser usadas para compor novos atributos mais poderosos

Dados faltantes – 1

- Solução mais simples: eliminação dos dados com valores faltantes
- Mesmo com grandes quantidades de dados, o subconjunto de dados completos pode ser bem pequeno
- Para manter todas as instâncias, fornecem-se manualmente valores razoáveis, prováveis ou esperados, em função da experiência

Dados faltantes – 2

- Substituição automática de valores faltantes por constantes tais como
  - Substituir todos os valores faltantes por uma única constante global (altamente dependente da aplicação)
  - Substituir o valor faltante pelo valor médio do atributo
  - Substituir o valor faltante pelo valor médio do atributo para aquela classe

Substituição pelo Valor Médio

Position	Original sample	Position 11 missing	Preserve mean as estimate	Preserve variance as estimate
1	0.0886	0.0886	0.0886	0.0886
2	0.0684	0.0684	0.0684	0.0684
3	0.3515	0.3515	0.3515	0.3515
4	0.9874	0.9874	0.9874	0.9874
5	0.4713	0.4713	0.4713	0.4713
6	0.6115	0.6115	0.6115	0.6115
7	0.2573	0.2573	0.2573	0.2573
8	0.2914	0.2914	0.2914	0.2914
9	0.1662	0.1662	0.1662	0.1662
10	0.4400	0.4400	0.4400	0.4400
11	0.6939	?	**0.3731**	**0.6622**
Mean	0.4023	0.3731	0.3731	0.3994
Standard deviation	0.2785	0.2753	0.2612	0.2753
Size of error in the estimate			0.3208	0.0317

Substituição mantendo o Desvio Padrão

Position	Original sample	Position 1 missing	Preserve mean as estimate	Preserve variance as estimate
1	0.0886	?	**0.4336**	**0.1479**
2	0.0684	0.0684	0.0684	0.0684
3	0.3515	0.3515	0.3515	0.3515
4	0.9874	0.9874	0.9874	0.9874
5	0.4713	0.4713	0.4713	0.4713
6	0.6115	0.6115	0.6115	0.6115
7	0.2573	0.2573	0.2573	0.2573
8	0.2914	0.2914	0.2914	0.2914
9	0.1662	0.1662	0.1662	0.1662
10	0.4400	0.4400	0.4400	0.4400
11	0.6939	0.6939	0.6939	0.6939
Mean	0.4025	0.4336	0.4336	0.4076
Standard deviation	0.2791	0.2723	0.2584	0.2723
Size of error in the estimate			0.3450	0.0593

Dados faltantes – 3

- Valor faltante = "*don't care*"; a instância com valor faltante é substituída por novas instâncias, em que o valor faltante é substituído pelos possíveis valores do domínio

- Método mais popular: gerar um modelo preditivo para gerar cada um dos valores faltantes de cada um dos atributos

Análise de outliers – 1

- Detecção de outliers é o processo de seleção de *k* instâncias de um total de *n* que sejam consideradas inconsistentes com os dados restantes
  - A técnica mais simples está baseada na *estatística*. Neste caso, determinam-se a média e a variância da amostra e descartam-se as instâncias que estão além de um certo valor limite da variância

Análise de outliers – 2

- *Detecção de outliers baseada na distância*: s_i é um outlier se pelo menos uma fração *p* das instâncias está a uma distância maior que *d* (não tem vizinhos suficientes a uma certa distância)

- *Técnicas baseadas em desvio*: define o menor subconjunto de instâncias cuja remoção resulta na maior redução da função de dissimilaridade do conjunto residual

Redução de Dados

- Há um potencial de resultados melhores da mineração com grandes bases de dados, mas não há garantia que o conhecimento obtido seja melhor

- Um subconjunto dos dados preparados e pré-processados
  - pode ser obtido em um tempo razoável?
  - pode ser descartado sem afetar a qualidade dos resultados?

Operações sobre os Dados

- Operações básicas no processo de redução de dados:
  - deleção de colunas (atributos)
  - deleção de linhas (instâncias)
  - deleção de valores (de um atributo)

- Preservam as características dos dados

- Outras operações de redução de dados tornam os novos dados *não-reconhecíveis*

Dimensões dos Dados

- Redução de dados não reduz necessariamente a qualidade dos resultados (às vezes melhora)

- Parâmetros de comparação (redução):
  - Tempo de cálculo
  - Precisão preditiva ou descritiva
  - Simplicidade de representação do modelo

- Não existe um método de redução de dados que seja melhor em todos os casos

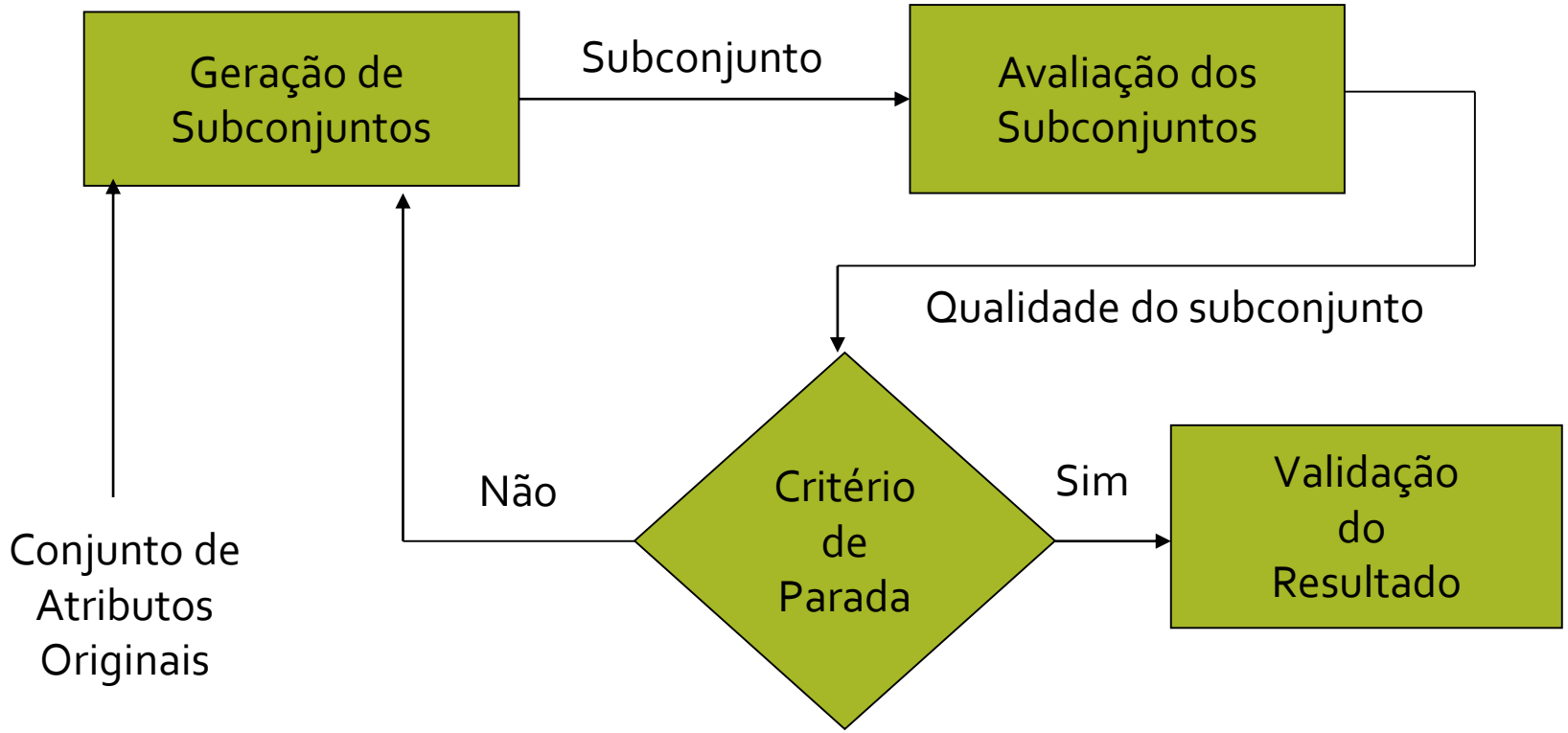
Redução de Atributos

- O processo de redução de dados resulta em:
  - menos dados: o algoritmo aprende mais rápido
  - maior precisão: o modelo generaliza melhor
  - resultados mais simples: a interpretação e o uso do modelo é mais fácil
  - menos atributos: na próxima coleta de dados a exigência é menor

Seleção de Características - Conceitos

- Escolher um subconjunto com *M* dos *N* atributos originais
- É um problema de busca
- Objetivos
  - Reduzir a dimensionalidade do espaço
  - Acelerar o algoritmo de aprendizagem
  - Melhorar a precisão da previsão
  - Melhorar a compreensibilidade

Seleção de Características – Passos Básicos



Seleção de Características – Procedimento de Busca

- Com N atributos originais, tem-se 2^N subconjuntos candidatos $\rightarrow$ busca exaustiva inviável

- Critérios de parada:
  - Número pré-definido de atributos
  - Número pré-definido de iterações
  - Subconjunto ótimo de acordo com o critério
  - Adição ou retirada de atributo não altera o melhor subconjunto

Seleção de Características – Geração de Subconjuntos

- Ponto de início
  - Conjunto vazio – geração avante
  - Conjunto completo – geração para trás
  - Conjunto aleatório – geração aleatória

- Estratégia de busca
  - Completa – baseada em propriedades
  - Heurística
  - Não determinista

Seleção de Características – Critério de Avaliação

- Filtro("filter)
  - Critério independente sem o envolvimento do algoritmo de aprendizagem
  - Exemplos: medidas de distância, de informação, de dependência, de consistência

- Envelope("wrapper")
  - Uso da avaliação do desempenho do algoritmo de aprendizagem quando aplicado ao subconjunto de atributos selecionado

Discretização de Atributos

- Consiste na transformação de atributos numéricos em atributos categóricos

- Existem duas possibilidades

  - Discretização local: usa informações das proximidades para determinar os pontos de corte

  - Discretização global: utiliza todos os valores que o atributo assume para estipular os pontos de corte

Discretização Não-Supervisionada

- Discretização não-supervisionada gera intervalos sem utilizar a informação da classe, e é a única possibilidade na tarefa de agrupamento

- Duas estratégias principais:
  - Intervalos de mesmo tamanho
  - Intervalos de mesma frequência

- Inferior a esquemas supervisionados em tarefas de classificação

Discretização Supervisionada

- Os intervalos são determinados em função dos valores do atributo e da classe corresponde a cada valor

- Diversos métodos:
  - Baseados em entropia
  - Baseados em programação quadrática