

Aula 03: Pré- processamento de Dados

PPGGCO – MINTER – INDAIAL

PROF. DR. IGOR DA PENHA NATAL

E-MAIL: IGOR.NATAL@UNICESUMAR.EDU.BR

Pré-Processamento

- Seleção
- Limpeza
- Codificação
- Enriquecimento
- Normalização
- Construção de Atributos
- Correção de Prevalência
- Partição do Conjunto de Dados

Os atributos/dados podem ser classificados em:

- Representação de seus valores (tipo de dados);
- Natureza da informação (tipo da variável) e as variáveis podem ser classificadas:
 - **Nominais ou categóricas**: São atributos que permitem rotular o conjunto de dados.
Ex: estado civil.
 - **Discretas**: Assemelham-se às variáveis nominais, mas os valores (estados) que elas podem assumir possuem um ordenamento.
Ex: dia da semana
 - **Contínuas**: São variáveis quantitativas, cujos valores possuem uma relação de ordem entre eles.
Ex: idade, renda...

Onde tudo começa...

- A etapa de pré-processamento também é denominada de ETL (Extract-transform-load).
- Tem por objetivo organizar os dados, padronizar a representação da informação e, quando possível, simplificar o processo de mineração de dados
- Muitas heurísticas podem ser aplicadas nesse processo.
- Segundo o Ralph Kimball, são 34 subsistemas que compõem o ETL e estima-se um consumo de 70% dos tempo e dos recursos na preparação e construção do data warehouse.
(Fonte: <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/etl-architecture-34-subsystems/>)
- A maioria dos métodos de mineração de dados pressupõe que os dados estejam organizados em uma única estrutura tabular.

Função: Seleção de Dados (1/5)

- A junção dos dados em uma única tabela pode ocorrer de duas formas:
 - Junção direta: todos os dados são unificados;
 - Junção Orientada: O especialista do domínio da aplicação, em parceria com especialista de KDD, escolhem os atributos e registros.
- O processo de escolha pode ser:
 - Redução de Dados Horizontal
 - Segmentação do Banco de Dados
 - Eliminação direta de casos
 - Amostra Aleatória
 - Amostragem Simples Sem Reposição
 - Amostragem Simples Com Reposição
 - Amostragem de Clusters
 - Amostragem estratificada

Função: Seleção de Dados (2/5)

- Agregação de informações
- Redução de Dados Vertical – entre as principais motivação para a aplicação da redução vertical dos dados, tempos:
 - Um conjunto de atributos bem selecionado pode conduzir a modelos de conhecimento mais concisos e com maior precisão;
 - Se o método de seleção for rápido, o tempo de processamento necessário para utilizá-lo e, em seguida, aplicar o algoritmo de mineração de dados em um subconjunto dos atributos, pode ser inferior ao tempo de processamento do algoritmo de DM;
 - A exclusão de atributos é muito mais significativa em termos de redução do tamanho do conjunto de dados que a exclusão de registros.

Existem duas abordagens para redução de dados verticais:

- Abordagem Independente de modelo (Filter)
- Abordagem dependente de modelo (Wrapper)

Função: Seleção de Dados (3/5)

Em qualquer das abordagens vistas, três estratégias clássicas e simples, para escolha do conjunto de atributos podem ser utilizadas (Ham & Kember, 1999):

- **Seleção Sequencial para Frente** (Forward Selection):

Inicia-se com o subconjunto de atributos vazio. Cada atributo é adicionado ao subconjunto de atributos candidatos, que é avaliado segundo alguma medida de qualidade. No final de cada iteração é adicionado o atributo que tenha maximizado a média de qualidade.

- **Seleção Sequencial para Trás** (Backward Selection):

Inicia-se com o subconjunto de atributos candidatos contendo todos os atributos, a cada iteração retira-se o atributo do conjunto que tenha minimizado a medida de qualidade considerada.

- **Combinação das Estratégias Anteriores:**

a cada passo do algoritmo, o melhor atributo é adicionado e o pior dentre os anteriormente escolhidos é removido.

- Outras técnicas: ID3, algoritmos genéticos

Função: Seleção de Dados (4/5)

- Eliminação direta de atributos – duas heurísticas podem ser utilizadas:
 - Eliminação dos atributos com valores constantes em todos os registros
 - Eliminação dos atributos identificadores de registros (chaves primárias, ou chaves naturais)
- Análise de Componentes Principais (*PCA – Principal Component Analysis*)
São procedimentos do PCA:
 - Normalização de todos os atributos;
 - PCA computa c vetores ortonormais que formam uma base para os dados de entrada normalizada;
 - Os componentes principais são ordenados em ordem decrescente de variância;
 - Elimina-se os componentes mais fracos (menor variância)

Função: Seleção de Dados (5/5)

- Redução de Valores
alternativa ao processo de redução horizontal (corte de atributos).
Prevê a redução dos valores distintos de um atributo;
- Redução de Valores Nominais
 - Identificação de hierarquia entre atributos;
 - Identificação de hierarquia entre valores;
- Redução de Valores Contínuos (ou Discretos)
 - Particionamento em células (Bins);
 - Redução de Valores pelas Medianas das Células (Bin Medians)
 - Redução de Valores pelas Médias das Células (Bin Means)
 - Redução de Valores pelos Limites das Células (Bin Boundaries)
 - Arredondamento de Valores
 - Agrupamento de Valores (Clusterização)

Função: Limpeza

- Chama-se de limpeza, os processos desenvolvidos para remover as inconsistências, incompletude ou ruídos nos dados.
- Evitar o efeito GIGO (garbage in, garbage out);
- Funções:
 - Limpeza de informações ausentes
 - Exclusão dos casos
 - Preenchimento manual de valores
 - Preenchimento com valores globais constantes
 - Preenchimento com medidas estatísticas
 - Preenchimento com métodos de mineração de dados (Ex: modelos bayesianos, árvores de decisão, probabilísticos, estatísticos etc).
 - Limpeza de Inconsistências
 - Exclusão de casos
 - Correções de erros
 - Limpeza de valores não pertencentes ao domínio

Função: Codificação

- Codificação é o processo de mudar a representação da informação, normalmente para algum valor numérico, para atender as necessidades das técnicas de mineração.
- Codificação: Numérica – categórica
 - Mapeamento direto: substitui os valores numéricos por valores categóricos.
Ex: campo sexo: 1-M,0-F
 - Mapeamento em intervalos (Discretização) – a representação em intervalos pode ser obtida a partir de métodos que dividam o domínio de uma variável numérica em intervalos.
 - Divisão em intervalos com comprimentos definidos pelo usuário
 - Divisão em intervalos de igual comprimento
 - Divisão em intervalos por meio de clusterização

Função: Codificação

- Codificação: Categórica – numérica
 - Representação Binária Padrão (econômica)
 - Representação Binária 1-de-N (Ex: Casado – 1, Solteiro – 10, Viúvo – 100...)
 - Representação Binária por temperatura – utilizada quando existe graduação entre os conceitos ou rótulos (Ex: 1-fraco, 11-Regular, 111-Bom e 1111-ótimo).

Função: Enriquecimento

- Enriquecimento é o nome dado a técnica que tenta agregar mais valores aos dados existentes.

Adicionando informação ou sentido aos dados disponíveis.

- Pesquisas:
obtenção de mais detalhes dos dados nas fontes que deram origem aos dados disponíveis no banco de dados.
- Consultas a Bases de Dados Externas:
consiste em enriquecer os dados com a incorporação de outras bases de dados, proveniente de outros sistemas.

Função: Normalização dos Dados

- Esta operação consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como -1 a 1, ou de 0 a 1.
 - Normalização linear (interpolação linear) – fórmula: $A' = (A - \min) / (\max - \min)$
 - Normalização por desvio padrão (Z-Score ou Zero Mean): $A' = (A - X) / \sigma$
 - Normalização pela soma dos elementos: $A' = A / X$
 - Normalização pelo valor máximo dos elementos: $A' = A / \max$
 - Normalização por escala decimal: $A' = A / 10^i$

Função: Construção de Atributos

- Essa operação consiste em gerar novos atributos a partir de atributos existentes.
- Os casos mais simples (comum) são com atributos datas.

Ex:

```
data -> idade  
(year(sysdate)-year(data_nasc))
```

Função: Correção de Prevalência

- Essa operação é muitas vezes necessárias em tarefas de classificação. Consiste em corrigir um eventual desequilíbrio na distribuição de registros com determinadas características.

Ex: considere a existência de 1% de casos de inadimplência.

- O método de replicação aleatória de registros pode ser uma solução;
- Aplicação de matriz de custo pode resolver problemas de prevalência. Consiste em ter o peso do erro associado a classe que sejam menos numerosas.

Função: Partição do Conjunto

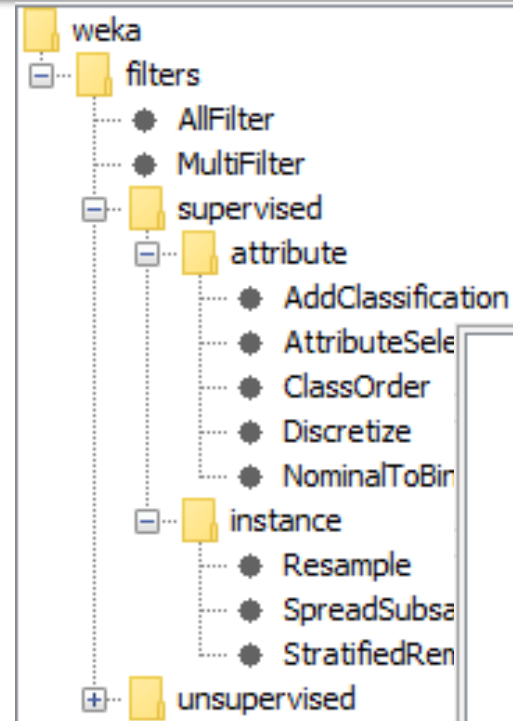
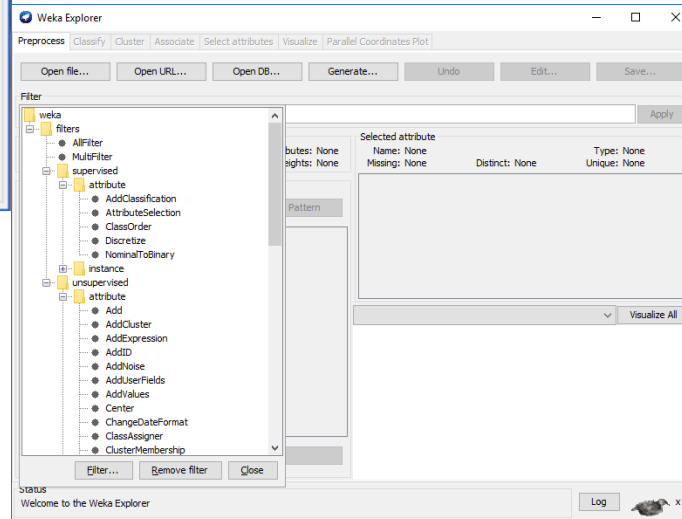
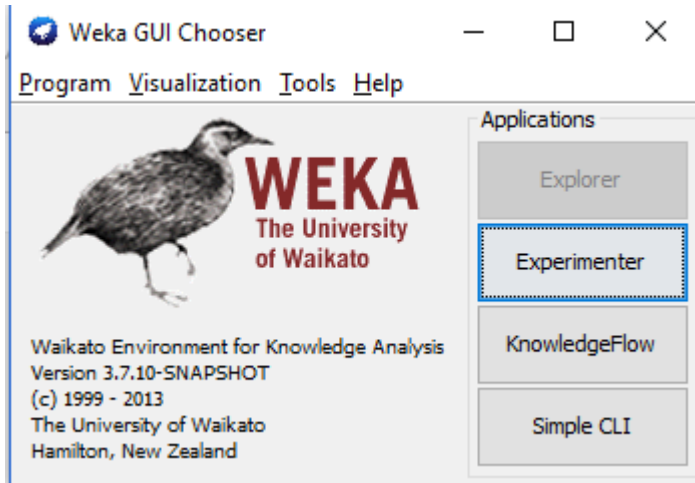
- Normalmente, para se validar um modelo, deve-se confrontá-lo com uma massa de dados que não tenha sido utilizada no aprendizado (problema: os algoritmos decoram os casos).
- Solução: dividir o conjunto de dados existentes, técnicas:
 - **Holdout**: divide o conjunto em um percentual p para treinamento e $(1-p)$ para teste. Normalmente $p > \frac{1}{2}$.
 - **K-Fold-CrossValidation** (Validação cruzada com k conjuntos): Divide o conjunto de dados em subconjuntos de N/k , e usa-se $k-1$ para teste e os demais são reunidos para o treinamento.
 - **Stratified K-Fold** (Validação cruzada com K conjuntos): divide o conjunto em subconjunto mutuamente exclusivos, preserva-se a proporção das classes.
 - **Leave-One-Out**: Caso particular do Stratified, onde cada subconjunto tem um elemento.
 - **Bootstrap** – O conjunto de treinamento é obtido pelo sorteio com reposição, o conjunto de teste é obtido com sorteio entre os casos não utilizados no primeiro conjunto.

Bibliografia

- GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.
- FAYYAD, U. M.; PLATETSKY-SHAPIRO, G.; SMYTH, P. *From Data Mining to Knowledge Discovery: An Overview*. **Knowledge Discovery and Data Mining**, Menlo Park: AAAI Press, 1996
- FAYYAD, U. M.; PLATETSKY-SHAPIRO, G.; SMYTH, P. The *KDD Process for Extracting Useful Knowledge from Volumes of Data*. **Communications of the ACM**, v. 39, 1996
- Weka
Canal:
<https://www.youtube.com/channel/UCXYXSGq6Oz21b43hpW2DCvw>
Introdução
<https://www.youtube.com/watch?v=Exe4Dc8FmiM>

No WEKA

- O WEKA possui:
 - 75 algoritmos para pré-processamento dos dados
 - 25 algoritmos para seleção de atributos
 - 13 algoritmos para filtro de instâncias (registros)



- NumericToNominal
- NumericTransform
- Obfuscate
- PartitionedMultiFilter
- PartitionMembership
- PKIDiscretize
- **PrincipalComponents**
- RandomProjection
- RandomSubset
- Remove
- RemoveByName
- RemoveType
- RemoveUseless
- RenameAttribute
- Reorder
- ReplaceMissingValues
- ReplaceMissingWithUserConstant
- SortLabels
- Standardize
- StringToNominal
- StringToWordVector
- SwapValues
- TimeSeriesDelta
- TimeSeriesTranslate