
UNIVASF
FUNDAÇÃO UNIVERSIDADE FEDERAL
DO VALE DO SÃO FRANCISCO

Descoberta de Conhecimento em Bancos de Dados - KDD

Professor: Rosalvo Ferreira de Oliveira Neto

Disciplina: Inteligência Artificial

Tópicos

1. Definições
2. Fases do Processo
3. Avaliação de Classificadores
4. Medidas de Desempenho

Descoberta de Conhecimento em Bancos de Dados - KDD

A descoberta de conhecimento em bancos de dados (Knowledge Discovery in Databases-KDD) é um processo que envolve desde a preparação da base de dados até a apresentação do conhecimento deles extraído pelas técnicas de mineração.

Tarefas Básicas

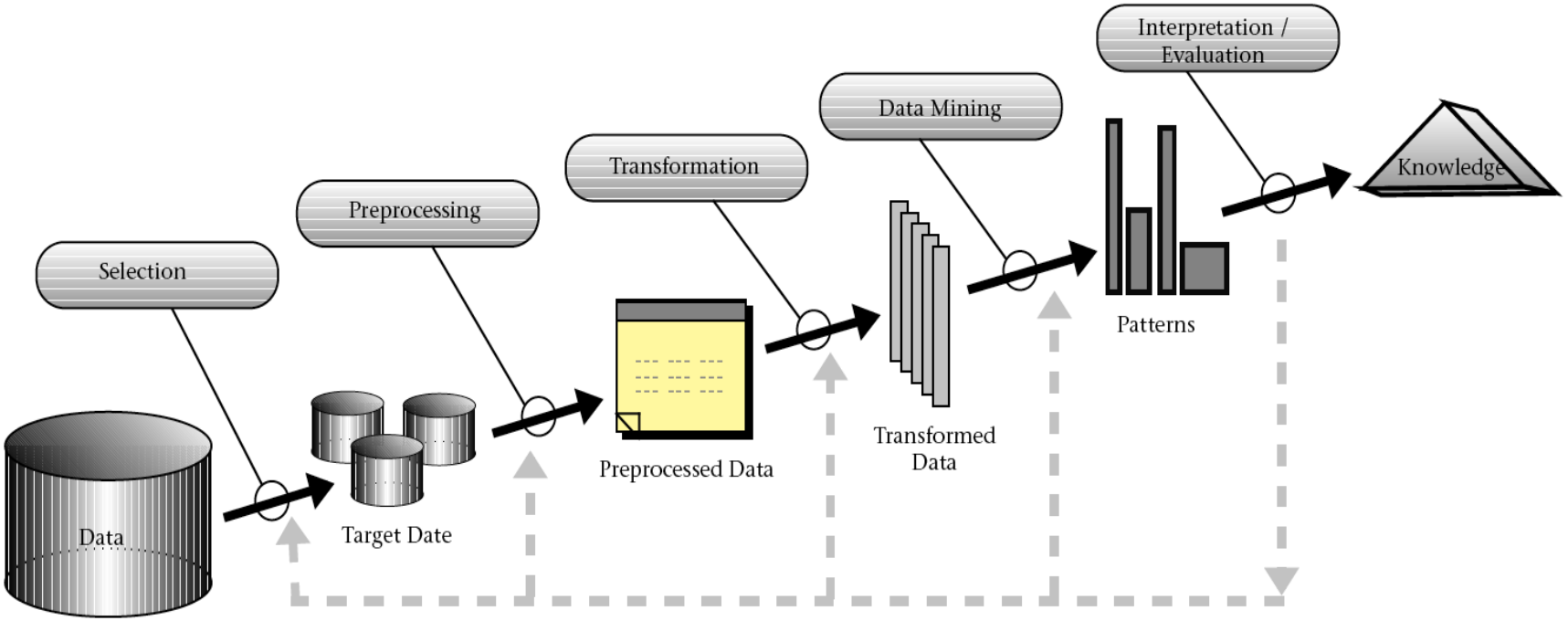
- ✓ Agrupamento
  - Identificação de grupos de indivíduos/registros que têm perfis semelhantes
- ✓ Regressão
  - Estimação de valores contínuos na resposta do sistema
- ✓ Classificação
  - Decisão do sistema categorizando cada indivíduo/registro em uma classe pré-definida
- ✓ Extração de regras de associação e de classificação
  - Apresentação de relações entre as variáveis de entrada e entre as variáveis de entrada e as respostas do sistema

Descoberta de Conhecimento em Bancos de Dados - KDD

A definição do termo Knowledge Discovery in Databases (KDD) foi introduzida por Fayyad et al. como parte de um processo ainda mais amplo de Data Mining.

"Knowledge Discovery in Databases (KDD) ou Descoberta do Conhecimento em Bases de Dados é um processo não trivial, iterativo, interativo e com múltiplos estágios que manipula e transforma os dados no intuito de descobrir padrões relevantes. Fayyad et al. identificaram cinco estágios do processo de KDD:"

Descoberta de Conhecimento em Bancos de Dados - KDD



Seleção

Seleção dos dados: também chamado de amostragem dos dados, é o processo que define quais serão os dados a serem minerados no projeto. Os dados podem ser selecionados das mais diversas fontes de dados, tais como: banco de dados relacional, arquivo texto legado, dentre outros.

Pré-processamento

Pré-processamento dos dados: é nesta fase que os dados são organizados e as inconsistências e integração são tratadas.

Mudança de granularidade, Tratamento de Missing Value e OutLiers

Transformação

Transformação dos dados: que consiste na transformação dos dados brutos em dados transformados para aplicação da técnica inteligente. Esta fase depende do algoritmo a ser aplicado na fase seguinte.

Data Mining

Data Mining: também conhecido como algoritmo de aprendizagem, esta fase aplica a técnica inteligente para extração do conhecimento. Na fase seguinte, é aplicado o algoritmo minerador, como por exemplo: redes neurais, árvores de decisão, análise de clustering, dentre outros.

Avaliação

Interpretação dos Resultados: Por fim, vem a fase de validação do conhecimento minerado, onde o especialista do domínio de aplicação é fundamental para homologação do conhecimento adquirido, pois nesta fase são validados todos os resultados obtidos no projeto

Avaliação de Classificadores

Avaliação

Existem poucos estudos analíticos sobre o comportamento de algoritmos de aprendizagem.

A análise de classificadores é fundamentalmente experimental.

Dimensões de análise:

- Taxa de erro
- Complexidade dos modelos
- Tempo de aprendizagem

…

Avaliação

Dois Problemas distintos:

- Dados um algoritmo e um conjunto de dados:

  Como estimar a taxa de erro do algoritmo nesse problema?

- Dados dois algoritmos e um conjunto de dados:

  A capacidade de generalização dos algoritmos é igual?

Avaliação

Qual o desempenho do modelo aprendido?

Erro no conjunto de treinamento não é um bom indicador em relação ao que vai ser observado no futuro.

Solução simples quando os dados são abundantes

dividir os dados em treinamento e teste.

Avaliação

Treinamento e teste

Medida natural de desempenho para problemas de classificação: taxa de erro

- Sucesso: a classe da instancia é prevista corretamente
- Erro: classe da instancia é prevista incorretamente
- Taxa de erro: proporção dos erros em relação ao conjunto de exemplos

- Erro de re-substituição: erro calculado a partir do conjunto de treinamento

- Erro de re-substituição é otimista!

Avaliação

É importante que os dados de teste não sejam usados de nenhuma maneira para construir o classificador

Alguns algoritmos de aprendizagem operam em dois estágios

- Estágio 1: construção da estrutura básica
- Estágio 2: otimização do ajuste dos parâmetros

Procedimento correto: usar 3 conjuntos: treinamento, validação e teste

- Validação: usado para otimizar os parâmetros

Avaliação

Estimação Holdout

O que fazer se os dados são limitados?

O método holdout reserva uma certa quantidade para teste e o restante para a aprendizagem usualmente, 1/3 para teste e 2/3 para treinamento

Problema: a amostra pode não ser representativa exemplo: uma classe pode estar ausente no conjunto de teste

Solução: amostragem estratificada: as classes são representadas com aproximadamente a mesma proporção tanto no teste como no treinamento

Avaliação

Holdout repetido

Estimação holdout pode ser realizada com mais confiança repetindo-se o processo com diferentes sub-amostras

- Em cada iteração, uma certa proporção é selecionada aleatoriamente para treino, com ou sem estratificação

- Uma taxa de erro global é calculada pela média das taxas de erro nas iterações

Problema: os diferentes conjuntos de teste não são mutuamente excludentes

Avaliação

Validação cruzada (validação cruzada k-fold**)**

- Os dados são divididos em k conjuntos de mesmo cardinal

- Cada subconjunto é usado como teste e o restante como treino

Validação cruzada evita conjuntos de teste com interseção

A taxa de erro global é a média das taxas de erro calculadas em cada etapa

Avaliação

Validação cruzada leave-one-out

- É uma forma particular de validação cruzada
- O número de folds é o número de exemplos
- O classificador é construído n vezes

- Não envolve sub-amostras aleatórias
- Computacionalmente custoso
- A estratificação não é possível

Avaliação

Bootstrap

Validação cruzada usa amostragem sem repetição

Bootstrap é um método de estimação que usa amostragem com reposição para formar o conjunto de treinamento

Retira-se uma amostra aleatória de tamanho n de um conjunto de n exemplos com reposição

Essa amostra é usada para o treinamento

os exemplos dos dados originais que não estão no conjunto de treino são usados como teste

É a melhor maneira quando o conjunto de dados é pequeno

Avaliação

Exemplos de medidas de desempenho

- Raiz do erro quadrático médio;
- A matriz “confusão”.

Avaliação

Raiz do erro quadrático médio

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

Avaliação

Matriz “confusão”

		Predicted class	
		Yes	No
Classe Atual	Yes	True positive	False negative
	No	False positive	True negative


UNIVASF
FUNDAÇÃO UNIVERSIDADE FEDERAL
DO VALE DO SÃO FRANCISCO