

USP

Data Mining (Ferramentas)

Solange Oliveira Rezende

USP São Carlos / ICMC
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
<http://www.icmc.usp.br>

© Solange Oliveira Rezende

USP

Técnicas e Ferramentas para DM

sor

USP

Ferramentas DM

Analista	Usuário Final	Programador
✓ DataMining Suite ✓ Darwin ✓ DataEngine ✓ DataMind ✓ Data Surveyor ✓ Discovery Server ✓ Intelligent Miner ✓ Omega ✓ SAS ✓ Syllogic DMT	✓ Alice ✓ BusinessMiner ✓ Clementine	✓ DataEngine ✓ Intelligent Miner ✓ Omega

sor

USP

Algumas Ferramentas para DM

✓ DataMining Suite ✓ Darwin ✓ DataEngine ✓ DataMind ✓ Data Surveyor ✓ Discovery Server ✓ Intelligent Miner ✓ Omega ✓ SAS ✓ Syllogic DMT ✓ Mineset ✓ Weka	✓ Alice ✓ BusinessMiner ✓ Clementine ✓ DataEngine ✓ Intelligent Miner ✓ Omega
---	--

sor

USP

DataMining Suite - Information Discovery

- ✓ **Características:**
 - ✧ Plataforma: Win3/95/NT.
 - ✧ Acesso a BD: Oracle, Sybase, Red Brick.
 - ✧ Vendedor de Corporate Vision (OLAP) e Icon Query (ferramenta para consultas).
- ✓ **Técnicas:**
 - ✧ Redes neurais.
 - ✧ Regras de indução.
- ✓ **Pontos Fortes:**
 - ✧ Pode trabalhar com grandes BD.
 - ✧ Poderosas técnicas de visualização.
- ✓ **Pontos Fracos:**
 - ✧ Limitado conjunto de técnicas.

sor

USP

Darwin - Thinking Machines

- ✓ **Características:**
 - ✧ Começo em 1983.
 - ✧ Plataformas: Unix, WinNT.
 - ✧ Acesso a BD.
- ✓ **Técnicas:**
 - ✧ Algoritmos Genéticos (StarGene).
 - ✧ K-nearest neighbor (StarMatch).
 - ✧ Redes Neurais (StarNet).
 - ✧ Regras de indução - CART (StarTree).
- ✓ **Pontos Fortes:**
 - ✧ Desenvolvido para trabalhar com grande conjuntos de dados.
 - ✧ Suporta plataformas paralelas.
 - ✧ Muitas técnicas.

sor

DataEngine - MIT

- ✓ **Características:**
 - ✧ Plataforma: Win3/95/NT.
 - ✧ DataEngine ADL = API..
- ✓ **Técnicas:**
 - ✧ Redes neurais.
 - ✧ Lógica fuzzy.
- ✓ **Pontos Fortes:**
 - ✧ Técnicas de visualização avançadas.
 - ✧ Parceria fechada com a comunidade acadêmica da Alemanha.
- ✓ **Pontos Fracos:**
 - ✧ Empresa semi-comercial.

DataMind - DataMind

- ✓ **Características:**
 - ✧ Três edições:
 - **DataCruncher:** Requer melhoramentos para que o PE e Solo consigam rodar sobre o servidor.
 - **Professional Edition:** Módulo cliente.
 - **Solo:** Módulo cliente integrado com Excel.
 - ✧ Plataformas DC: Unix, WinNT.
 - ✧ Acesso a BD: ODBC e arquivos planos.
- ✓ **Técnicas:**
 - ✧ Redes de agente (comparáveis com redes neurais).
- ✓ **Pontos Fortes:**
 - ✧ Integração com planilhas de cálculo.
 - ✧ Tecnologia avançada
 - ✧ Relativamente fácil de usar.

Data Surveyor - Data Distilleries

- ✓ **Características:**
 - ✧ Plataforma: Unix.
 - ✧ Acesso a BD: SQL.
- ✓ **Técnicas:**
 - ✧ Árvores de decisão.
 - ✧ Regras de indução.
 - ✧ Regras de associação.
- ✓ **Pontos Fortes:**
 - ✧ Pode construir regras estatísticas.
 - ✧ Motor extensível.
- ✓ **Pontos Fracos:**
 - ✧ Versão Windows ainda não disponível.

Intelligent Miner - IBM

- ✓ **Características:**
 - ✧ Plataforma cliente: Win95/NT, AIX, OS2.
 - ✧ Plataforma Servidor: AIX, AS/400, MVS.
 - ✧ Acesso a BD: DB2, DataJoiner, arquivos planos.
- ✓ **Técnicas:**
 - ✧ Redes neurais (Kohonen, back propagation).
 - ✧ Regras de indução.
 - ✧ Regras de associação.
 - ✧ Análise de padrões no tempo.
- ✓ **Pontos Fortes:**
 - ✧ Pode trabalhar com grandes BD.
 - ✧ Muitas características de pré-processamento (sobre DB2).
- ✓ **Pontos Fracos:**
 - ✧ Limitado em visualização.
 - ✧ Limitada interatividade.
 - ✧ Só disponível para análise estatística.

SAS - SAS

- ✓ **Características:**
 - ✧ Plataforma: quase todas.
 - ✧ Acesso a BD: quase todas via gateway.
- ✓ **Técnicas:**
 - ✧ Redes neurais.
 - ✧ Algoritmos genéticos
- ✓ **Pontos Fortes:**
 - ✧ Pode trabalhar com grandes BD.
 - ✧ Conjunto integrado de ferramentas SAS, módulos e aplicações.

Syllogic Data Mining Tool

- ✓ **Características:**
 - ✧ Plataforma: Unix.
 - ✧ Antigamente: Solomon.
- ✓ **Técnicas:**
 - ✧ Regras de indução.
 - ✧ Regras de associação.
- ✓ **Pontos Fortes:**
 - ✧ Pode trabalhar com grandes BD.
 - ✧ Muitas técnicas.
 - ✧ Fácil de usar.
 - ✧ Parceria com Tandem - NonStop SQL.
- ✓ **Pontos Fracos:**
 - ✧ Marca de produto recente.

Alice - Isoft

- ✓ **Características:**
 - ✦ Plataforma: Win3/95/NT.
 - ✦ Acesso a BD: ODBC e arquivos planos.
 - ✦ A ferramenta mais poderosa é AC₂.
- ✓ **Técnicas:**
 - ✦ Regras de indução.
- ✓ **Pontos Fortes:**
 - ✦ Desenvolvimento interativo de árvores de decisão.
 - ✦ Suporta um modo interativo.
 - ✦ Muito barato.
 - ✦ Recomendável para especialistas não estatísticos.
- ✓ **Pontos Fracos:**
 - ✦ Limitado conjunto de técnicas.

BussinessMiner - Bussines Objects

- ✓ **Características:**
 - ✦ Plataforma: Win3/95/NT.
 - ✦ Internamente o Alice é usado.
- ✓ **Pontos Fortes:**
 - ✦ Integrado com BussinesObjects (ferramenta query/OLAP).
 - ✦ Barato (\$500 dólares).
 - ✦ Solução integrada para consultas, OLAP e DM.
- ✓ **Pontos Fracos:**
 - ✦ Classificações geradas com BM não podem ser importadas dentro de BO.

Clementine - Integral Solutions

- ✓ **Características:**
 - ✦ Começo em 1989.
 - ✦ Plataformas: Unix, VAX/VMS, WinNT.
 - ✦ Acesso a BD.
 - ✦ Parceria com Platinum e NCR.
- ✓ **Técnicas:**
 - ✦ Visualização.
 - ✦ Redes Neurais.
 - ✦ Regras de indução.
- ✓ **Pontos Fortes:**
 - ✦ Usuários usam uma metáfora de programação visual para criar os modelos.
 - ✦ Muitas ferramentas para pré-processamento.
 - ✦ Existem muitos clientes.
 - ✦ Recomendável para especialistas não estatísticos.

Marksman - HNC Software

- ✓ **Características:**
 - ✦ Plataforma: WinNT.
 - ✦ Acesso a BD: Sybase e arquivos planos.
- ✓ **Técnicas:**
 - ✦ Redes neurais.
- ✓ **Pontos Fortes:**
 - ✦ Pode trabalhar com grandes BD.
- ✓ **Pontos Fracos:**
 - ✦ Não portátil, vem junto com o hardware.
 - ✦ Não é de uso geral.

Mineset™

- ✓ Produzido pela Silicon Graphics™
- ✓ Constituído de algoritmos de Aprendizado de Máquina (MLC++) e de Ferramentas de Visualização Multidimensional dos Dados
- ✓ Principais Ferramentas de Visualização:
 - ✦ Visualizador de Árvores e de Regras de Decisão;
 - ✦ Visualizador Estatístico;
 - ✦ **Visualizador Bidimensional e Tridimensional;**
- ✓ Ferramentas de visualização podem auxiliar usuários a encontrarem padrões ou relações não-observados
- ✓ Roda também em Windows

Weka

- ✓ Biblioteca de algoritmos de Aprendizado de Máquina implementados em Java
- ✓ Paradigmas implementados: estatístico, *instance-based* e proposicional. Por exemplo: J48 (versão Java do C4.5), ID3, Naive Bayes, entre outros
- ✓ Os algoritmos podem ser utilizados isoladamente ou por uma aplicação Java escrita pelo usuário
- ✓ Pode ser conseguida em:
 - <http://www.cs.waikato.ac.nz/~ml/weka>

Outras ferramentas DM

- ✓ **KnowledgeSeeker - Angoss**
 - ✧ Plataforma: Unix.
 - ✧ Técnicas: Regras de indução.
- ✓ **Omega - CAP**
 - ✧ Plataforma: Win.
 - ✧ Técnicas: Algoritmos genéticos.
 - ✧ É orientado a modelos financeiros.
- ✓ **SPSS CHAID**
 - ✧ Técnicas: Regras de indução.
- ✓ **4Thought**
 - ✧ Técnicas: Redes neurais.