# Metodologias de *Data Mining*

Armando B. Mendes
amendes@notes.uac.pt

Açoriano Oriental (Bits & Bytes) de 23 de Junho de 2007.

Certamente *data mining* não são simples consultas a bases de dados ou tecnologias de integração em cubos de dados, nem mesmo um conjunto de algoritmos como redes neuronais, algoritmos genéticos, métodos estatísticos, *etc.*. A prospecção de dados é um processo ou uma metodologia para a descoberta de conhecimento que pode envolver todas os conceitos anteriores.

O modelo processual CRISP-DM (*CRoss Industry Standard Process for Data Mining*) não é proprietário e pretende ser independente do sector e das aplicações onde é utilizado. Esta metodologia tem sido validada com projectos de grande dimensão e resulta do trabalho conjunto de um utilizador intensivo, um fabricante de *software* e um consultor especialista em armazéns de dados. Actualmente, encontra-se em revisão que pode ser participada por todos (http://www.crisp-dm.org/).

Na versão 1.0, recomendam-se seis fases. Começa-se por compreender o problema e o contexto onde surge, incluindo a definição de objectivos e um plano de acção. Esta fase gera conhecimento de domínio que será utilizado durante todo o processo. A fase dois compreende a recolha, a integração, exploração e compreensão dos dados. A preparação e pré-processamento incluem tarefas de redução, transformação e limpeza de dados. Nestas últimas fases, tecnologias de integração em armazéns de dados e de formação de cubos são muito úteis.

Na fase de estimação ou aprendizagem de modelos, são utilizados algoritmos da estatística e da aprendizagem automática, como os atrás referidos. Na fase cinco, os resultados são validados, comparados, interpretados e confrontados com conhecimento de domínio, permitindo identificar conhecimento novo. Na última fase, a divulgação e implementação pode ser a simples escrita de um relatório, ou incluir a criação de modelos ou

mesmo uma aplicação integrada no sistema de informação. Em qualquer dos casos, pretende fazer chegar o conhecimento aos decisores. Apesar das fases bem definidas, o processo não é linear e apresenta imensos ciclos e retornos, mais coerente com uma espiral de modelação e extracção de conhecimento.

Tal como em todas as metodologias, a CRISP-DM não garante resultados, mas permite disciplinar o processo, perceber o papel dos vários intervenientes e alinhar os objectivos com o negócio.