Dolování dat

Ing. Roman Danel, Ph.D. **2010**

Co je to dolování dat (Data Mining)?

Data Mining (neboli dolování dat) je **proces hledání informací a znalostí ve velkém objemu dat**. Jedná se o nástroj používaný v rámci **Business Intelligence** (oblast analýzy dat sloužící jako podklady pro manažerské rozhodování).

Pro Data Mining se využívají technologie rozpoznávání vzorů, statistické a matematické metody.

Definice Data Miningu

> ***Data Mining je proces výběru, prohledávání a modelování ve velkých objemech dat, sloužící k odhalení dříve neznámých vztahů mezi daty za účelem získání obchodní výhody.***
>
> **Fayyad**

Dolování dat je hledáním skrytých souvislostí, procesem výběru, prohledávání a modelování ve velkých objemech dat. Slouží k odhalení dříve neznámých vztahů mezi daty.

V literatuře a zdrojích na webu je určitá nejednotnost v používání pojmu dolování dat (Data Mining) a **dobývání znalostí z databází** (Konwledge Discovery in Databases, KDD). Někteří autoři tyto pojmy používají jako synonyma. Podle první mezinárodní konference KDD (Montreal, 1995) je dolování dat součástí KDD.

Definice KDD

> ***Knowledge Discovery In Databases (KDD) je proces netriviálního objevování implicitních, dopředu neznámých a potenciálně použitelných znalostí v datech.***
>
> **Fayyad**

Oba pojmy tedy víceméně znamenají totéž, u KDD je považována jako důležitá i samotná příprava dat.

Dolování dat je úzce spojeno s pojmem **datový sklad**. Pro dolování je důležitá kvalita vstupních dat; datové sklady obvykle obsahují data, která jsou už v určité míře předzpracovaná a očištěná od chyb.

Při analýze pomocí dolování dat **často není dopředu známo, zda budou získány použitelné výsledky**. Rovněž **interpretace získaných výsledků** je považována za nejnáročnější fázi dobývání znalostí.

Software pro KDD se v literatuře označuje **Decission Support Systém** (DSS). Nejznámější softwarové produkty v této oblasti jsou Enterprise Miner (od firmy SAS) a Clementine (firma SPSS). Existují i nekomerční produkty jako Weka, Orange, Tanagra atd. V České republice se systémy DSS zabývá například firma Adastra nebo projekt Lisp-Miner (VŠE Praha)

Metody dolování dat

Dnes užívanými metodami dolování dat jsou například:

- regresní metody (lineární regresní analýza, nelineární regresní analýza, neuronové sítě)
- klasifikace (diskriminační analýza, logistická regresní analýza, rozhodovací stromy, neuronové sítě),
- segmentace – shlukování (shluková analýza, genetické algoritmy, neuronové shlukování – Kohonenovy mapy)
- analýza vztahů (asociační algoritmus pro odvozování pravidel typu „ if X then Y“)
- predikce v časových řadách (Boxova-Jenkinsonova metoda, neuronové sítě, autoregresní modely, ARIMA)
- detekce odchylek

Modely dolování dat

Deskriptivní model – popisuje nalezené vzory a vztahy v datech, které mohou ovlivnit rozhodování (Př. Analýza prodeje zboží v supermarketu, na jejímž základě je pak umístěno zboží v regálech).

Prediktivní model – umožňuje předvídat budoucí hodnoty atributů na základě nalezených vzorů v datech (Př. Analýza zákazníků, u kterých je vysoká pravděpodobnost, že budou reagovat na písemnou reklamní nabídku…)

Co je overfitting (přeučení)? Čím je způsobeno a jak mu zabránit?

- Přeučení modelu u data miningu
- Naučený model je příliš svázán s trénovacími daty
- Přesnost modelu je vysoká na trénovacích datech, ale nízká na nových datech
- Jak mu zabránit
    1. Rozdělení trénovacích dat (učení – test)
    2. Rozhodovací stromy – prořezávání, menší hloubka stromu
        1. Některé algoritmy ukončí včas generování stromu (prepruning)
        2. Většina nejdříve vygeneruje strom a pak ho ořeže (postpruning)
        3. Prořezávání zvyšuje chybu na učící množině, ale doufáme, že na reálných datech chybu zmenší

Metodologie dolování dat

Cílem metodologií je poskytnout uživatelům jednotný rámec pro řešení různých úloh z oblasti dobývání znalostí. Tyto metodologie umožňují sdílet a přenášet zkušenosti z úspěšných projektů.

Nejpoužívanější metodologie jsou: SEMMA (SAS), 5A (SPSS) a CRISP-DM.

CRISP-DM

CRISP (CRoss Industry Standard Proces for Data Mining) je souhrnná metodologie dobývání znalostí z databází, umožňuje provádět rozsáhlé projekty dobývání rychleji, efektivněji a méně nákladně prostřednictvím osvědčených postupů.

Základní etapy procesu dobývání jsou:

- Co řešit (Business understanding) – porozumění problematice, formulování úlohy
- Kde vzít data (Data understanding)
- Jak data připravit (Data preparation)
- Jak data analyzovat (Data modelling)
- Co jsme zjistili (Evaluation) – porozumění výsledkům
- Jak výsledky využít (Deployment)

SEMMA

Podle metodologie SEMMA spočívá proces dobývání v těchto krocích:

- **Sample -** vybírání vhodných objektů
- **Explore -** vizuální explorace a redukce dat
- **Modify -** seskupování objektů a hodnot atributů, datové transformace
- **Model -** analýza dat
- **Assess -** porovnání modelů a interpretace

5A

- **Assess** - posouzení potřeb projektu
- **Access** - shromáždění potřebných dat
- **Analyze** - provedení analýz
- **Act** - přeměna znalostí na akční znalosti
- **Automate** - převedení výsledků analýzy do praxe