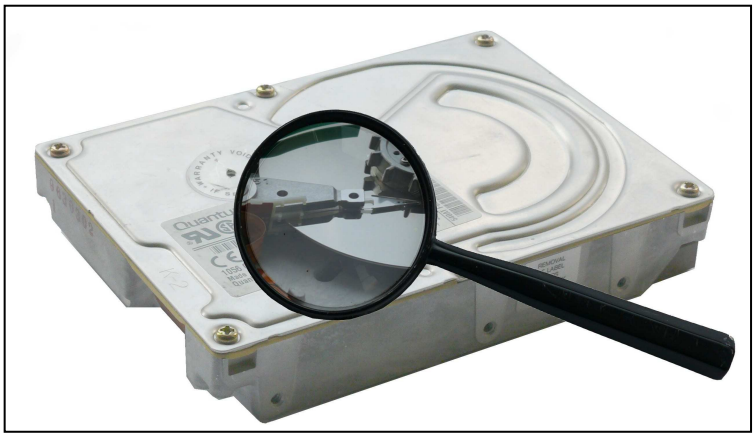
Datové dolování – Data mining

Datovým dolováním se nazývá proces, kdy se pomocí různých matematických metod pokoušíme nalézt v datech nějaké vnitřní, zatím neznámé, souvislosti. Obvykle se užívá na data shromážděná v předchozích obdobích a cílem je získání odpovědi na nějakou otázku související s daty. Příkladem může být například odhalení smluv, která klienti uzavřeli s cílem nějakého podvodu nebo určení pravděpodobného vývoje nějakého subjektu v budoucnosti. Snahou je dosáhnout toho, aby datové dolování dokázalo rozhodovat podobně jako zkušený expert z daného oboru, který se podívá na údaje o nové smlouvě a zjistí, že mu připadá podezřelá. Podobnou intuici, která je dána zkušenostmi, si může datové dolování vytvořit na základě učení ze starších dat.



Při procesu řešení se nejdříve data zpracují tak, že se nečíselné údaje převedou na číselné, pak se všechny údaje pomocí běžných statistických metod analyzují a vyřazují se údaje, které s hledanou odpovědí nemají žádný vztah. V dalším kroku se pak vyřadí data, která jsou poškozená nebo jsou zřetelně nesprávná. Po této předpřípravě dat, následuje opakované hledání souvislostí mezi daty, kdy se prověřují vazby mezi jednotlivými údaji a hledanou odpovědí. Po nalezení vhodných vztahů se zpracuje výsledné řešení a na závěr se ověří jeho kvalita (což vlastně znamená pravděpodobnost správné odpovědi). V přeneseném smyslu se matematický systém učí ze zkušeností a pak se testují jeho znalosti. Pokud jsou znalosti nedostatečné, pak se systém učí ze stejných dat jiným způsobem, dokud není nalezen nejlepší postup.
Paradoxem metody datového dolování je skutečnost, že výsledkem bývá relativně jednoduchý matematický vztah či snadný postup získání hledané odpovědi. Tento jednoduchý výsledek je však dán složitým a mnohonásobně opakovaným matematickým testováním zkoumaných dat.
Výsledek lze často vyjádřit formou stromu postupného rozhodování, odpovídající například klíči k určování hornin nebo rostlin. Další častou formou je nalezení matematického vztahu mezi údaji ve formě nějakého vzorce. A třetím způsobem je popis poměrů neuronové sítě (váhy jednotlivých parametrů a postup jejich slučování).

Omezení použitelnosti a ekonomický přínos
Celý postup datového dolování má několik omezení. Může se stát, že hledaná odpověď nijak s existujícími údaji nesouvisí. Příkladem budiž Švejk, který popsal dům a zeptal se, v kterém roce zemřela domovníkovi babička. Opačným extrémem může být závislost na příliš mnoha údajích. Příkladem může být předpovídaní počasí nebo budoucího pohybu kurzů, kdy nalezení odpovědi s vysokou pravděpodobností jistoty převyšuje dosavadní rychlost a kapacitu výpočetní techniky. Omezení jsou také dána chybami v datech (například některé údaje jsou neznámé či jsou nepřesné).
Dalším omezením je obvykle vždy existující nejistota o správném výsledku, kterou určujeme jako míru pravděpodobnosti správné odpovědi. Například stanovíme podezřelé smlouvy s 50% chybou. To znamená, že pokud se na základě výsledku datového dolování určí například 100 podezřelých smluv, tak pravděpodobně jen polovina z nich může být uzavřena podvodně. Ale pokud je výsledek dán testováním několika tisíc smluv, tak je uvedená míra pravděpodobnosti dostačující.
Ekonomický efekt řešení získaného datovým dolováním je zřejmý. Pokud například ztráty ze špatných smluv sníží využití této metody, třeba na polovinu, pak lze reálný přínos snadno vyčíslit.

Zjednodušený příklad datového dolování

Postup datového dolování lze předvést na jednoduchém příkladě, kterým může být hledání kritérií na oslovení tanečníků do soutěže Star Dance. Potřebujeme kandidáty, kteří by zaujali a přilákali dostatečný počet diváků pro inzerenty. Nejdříve se sestaví data, ze kterých se systém bude učit. Zařadí se do nich dřívější účinkující v podobných soutěžích (osoba je označena jako "VHODNÁ)" a zařadí se tam též údaje náhodně vybraných jiných osob (označí se jako „NEVHODNÁ"). Údaji budou věk, výška, váha. V tomto jednoduchém případě pomocí datového dolování zjistíme, že vhodný kandidát musí mít věk v rozmezí 20 až 35 let, výšku 175-185 cm a váhu nižší než výška osoby – 95. Výsledek je tedy vyjádřen rozhodovacím stromem o pouhých čtyřech řádcích:

1. Pokud je věk osoby mimo rozmezí 20 až 35 let, tak je osoba „NEVHODNÁ".
2. Pokud je výška mimo rozmezí 175 cm až 185 cm, tak je osoba „NEVHODNÁ".
3. Pokud je váha v kilogramech vyšší než výška v centimetrech - 95, tak je osoba „NEVHODNÁ".
4. Zbylé osoby jsou v kategorii „VHODNÁ".

Na základě takto určeného výsledku je zřejmé, které kandidáty je vhodné oslovit a které lze vynechat.
Prezentovaný příklad je velmi jednoduchý a hledaná klasifikace tvoří jen jednu skupinu. V praxi obvykle výsledky tvoří více skupin, například jiné parametry jsou „správné" pro malé podniky a jiné pro velké (zde by se rozhodovací strom rozdělil dvou částí dle typu podniku). Rovněž údaje nebylo třeba nijak upravovat a používají se v rozhodovacím stromu přímo.

Ostatní aplikace

Dalším častým úkolem datového dolování bývá nalezení shodných prvků. Často se využívá například v maloobchodní sféře, kdy po zjištění, že nejčastěji se spolu kupuje pečivo a mléčné výrobky, se tyto produkty umísťují co nejdál od sebe a akční ceny se vypisují střídavě jen na jeden z nich. Ve finančnictví lze tento způsob datového dolování využít pro seskupování produktů do balíčků, tvorbu akčních nabídek, atd.
Datové dolování se využívá i při získávání znalostí z časových řad, kdy se hledají významné vzorky, ze kterých lze předpovědět další vývoj. Například předpověď poruchy strojů nebo předpověď chování davu. Obecně se při hledání řešení pomocí datového dolování využívá i integrace či derivace vstupních údajů dle času, neboť výsledné řešení je často založeno i na tomto údaji (např. pokud hodnota parametru A za poslední rok stoupla o 50%, tak...).
Datové dolování je nadstavbou statistických metod. Pomocí vhodného zobrazení a metodou pokus-omyl může schopný analytik odhalit vazbu mezi několika málo údaji. Datové dolování je však schopné odhalit vazbu mezi velkým počtem údajů a tyto vazby zpracovat tak, aby výsledkem bylo použitelné řešení.

Forma smluvní spolupráce

Spolupráce mezi objednatelem a řešitelem datového dolování probíhá na základě uzavřené smlouvy. Objednatel určí otázku (nebo více otázek), na kterou požaduje odpověď a dodá data (či vybranou podmnožinu dat), která se mohou vztahovat k dané otázce. Data jsou v anonymní podobě, aby nebylo možné zjistit údaje, které si objednatel nepřeje (nepředávají se adresy, identifikační čísla, názvy, atd.). Pro stanovení řešení je vhodné, když je řešitel obeznámen s typem jednotlivých údajů (např. obrat, jmění, atd.), v případě úplné anonymity dat (údaje jsou označeny A, B, atd.) je řešení ztíženo, neboť nelze použít odhadnutelné logické předpoklady a řešení je časově mnohem náročnější.
Výsledkem prvotní analýzy dat je určení, zda existuje nějaký vztah mezi daty a položenou otázkou nebo v opačném případě, zda na položenou otázku nelze odpovědět dle zkušenosti z dodaných dat. Pokud nejsou nalezeny žádné použitelné vazby, objednatel se pokusí dodat další data (když jsou k dispozici). V případě, že je zjištěna závislost, pak probíhá navazující řešení na bázi datového dolování, kdy se systém na datech opakovaně „učí" hledat správnou odpověď na danou otázku, dokud není nalezena vhodná úprava vstupních údajů a učební postup vedoucí k nepřesnější odpovědi. Na základě úspěšného datového dolování předá řešitel objednateli jedno či více nalezených řešení k vlastnímu ověření na svých datech, kterým by měla být potvrzena správnost nalezeného řešení a jeho relevantnost.