

Data mining

jako prostředek konkurenční výhody

Miloš Uldrich

Každá firma si dnes eviduje velké množství údajů týkajících se rozličných procesů probíhajících v podniku. Schraňuje informace o transakcích, fakturách, svých zákaznících apod. Objem ukládaných dat neustále roste, a proto by byla škoda využívat je pouze k běžnému reportování.

Co je data mining

Data mining je dnes zcela určitě nejrychleji rostoucím segmentem business intelligence. S jeho pomocí se snažíme z ukládaných dat získat složitější a užitečnější informace než jen grafy a základní přehledy.

Ze statistického úhlu pohledu se jedná o hledání korelací, tedy vyšetřování vzájemných vztahů nebo vzorů v datech. Smyslem je tedy analyzovat datové závislosti, určit trendy, a pokud to typ dat umožňuje, předpovědět budoucí vývoj.

Informace pro všechny

Data mining pomáhá hledat odpovědi na různé otázky. Za praktický příklad může posloužit třeba snaha o zefektivnění firemního e-shopu. Pomocí uložených údajů o přístupech lze určit typické problémy při procházení webu, vysledovat závislosti mezi objemem prodeje a datem uskutečnění, určit komponenty, které zákazníci za daný čas nakoupili a v jakém pořadí (cross-selling). Vznikne tak analýza nákupního chování spotřebitele. Na základě ní lze navýšit zakázku doporučením souvisejícího produktu. Příkladem je výhodné balení fotoaparátu a čistící sady na objektiv, kdy čistící sada využívá většího objemu prodeje fotografických přístrojů. Tvorba efektivnějších produktových balíčků, a vůbec celá marketingová strategie je jen jednou oblastí použití. Využití data miningu lze doporučit také v oblasti detekce potenciálních podvodů, kontroly kvality výroby v průmyslové sféře, v různých predikcích a v mnoha dalších oborech včetně chemických laboratoří a nemocnic.

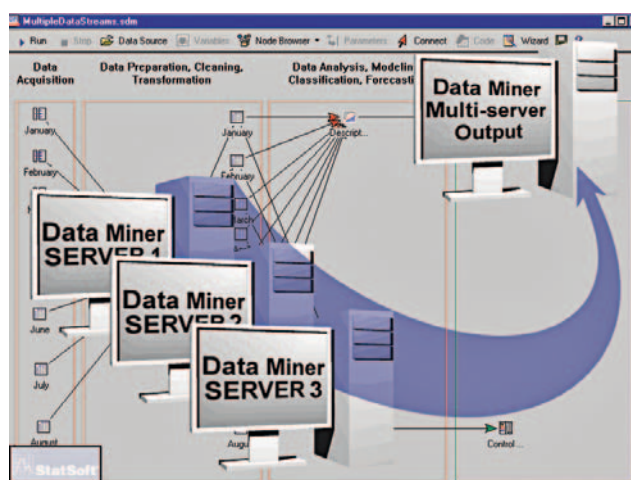
Z uvedeného vyplývá, že oblast aplikace je velice široká, a je zřejmé, že data mining již není výsadou pouze velkých společností, ale jak ukazuje praxe, tyto postupy mají nemalý potenciál i v menších firmách.

Požadavky na data

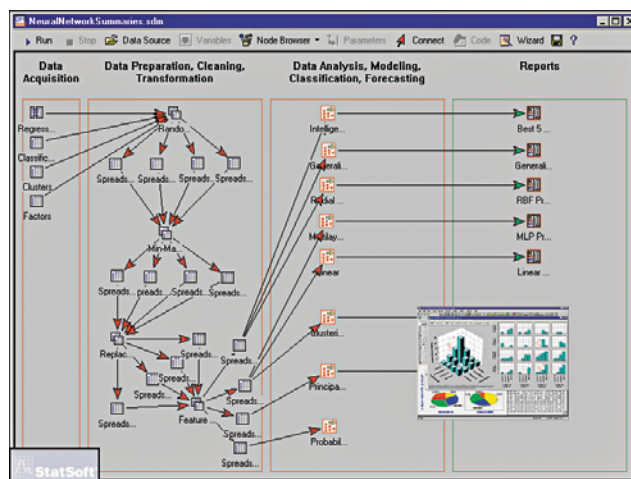
Jak nastiňují výše uvedené definice, z dat je sice možné získat cenné závislosti a ty přeměnit na údaje pro podporu tvorby další obchodní strategie, ale tyto vazby musí v datech skutečně existovat.

Z náhodných dat obvykle nic nesestavíme, případně může být výsledek tak triviální, že je pro praxi irrelevantní. Vždy je potřeba vhodně navrhnout vstupy, případně modifikovat stávající, například jejich sloučením do intervalu (u dotazníkových odpovědí se často omezuje počet diskretních hodnot atd.), a eliminovat chyby, které se mohou v primárních datech vyskytovat – například zákazník uskutečnil nákup před datem svého narození, nakoupil zboží v prodejně, která ještě nebyla otevřena apod. Úprava datového skladu do nezbytné formy často vyžaduje různé formy agregace. Důležité je také neopomenout zachytit všechny relevantní údaje k dané

problematice. Nelze totiž objektivně zkoumat závislost prodeje piva bez zahrnutí vlivu počasí či závislost prodeje energetického nápoje bez vlivu měsíce, protože lze předpokládat, že v různých obdobích bude určitá skupina (např. studenti) konzumovat tento nápoj ve větším objemu. Návrh úpravy datového skladu tak, aby uložená data měla co největší hodnotu, by měl být pro každou firmu poskytující software z oblasti data miningu samozřejmostí.



Obr.1: Schéma možnosti zpracování analýz STATISTICA Data Miner



Obr.2: Ukázka prostředí STATISTICA Data Miner

Maximální možná kvantifikace vztahů v datech musí být prioritou všech společností snažících se získat cenné informace, a být tak vždy o krok před konkurencí. ■

Autor je konzultant společnosti StatSoft CR s.r.o.
Foto: archiv StatSoft