Využití dataminingových metod v praxi

Lenka Blažková, Miloš Uldrich

Pojďme se podívat podrobněji na úlohy, které svou povahou nespadají do statistiky jako takové, a sice na postupy, jež se souhrnně označují jako prostředky pro data mining neboli dolování dat. Tyto metody umožňují objevit či zpřesnit dosud neznámé závislosti v datových souborech různého zaměření. Cíle analýzy jsou různé, od popisu chování vybrané skupiny zákazníků až po vytvoření spolehlivých pravidel, podle kterých budeme třídit žadatele o úvěry.

Jak využiji klasifikační úlohy?

Na základě klasifikace lze data, pokud jsou k tomu vhodná, řadit do určitých tříd a kategorií. Záměrem klasifikační úlohy je každý jednotlivý případ v souboru zařadit do konkrétní odhadnuté třídy. Tyto metody podávají velmi dobré výsledky například při vyhodnocování rizika poskytovaného úvěru (na základě záznamů si uživatele roztřídíme do skupin, jež reprezentují například stupeň rizikovosti klienta). Obecně je klasifikační úloha založena na výpočtu podmíněných pravděpodobností jednotlivých zařazení (Bayesův teorém), které jsou určeny na základě četnosti v datech z minulosti.

Co jsou to asociační pravidla

Asociační pravidla definují zajímavé vztahy mezi datovými položkami v databázi. Tato metoda se používá například k analýze spotřebního chování nakupujících, kdy jsou zkoumány jednotlivé položky nákupního koše, mezi kterými jsou potom hledány vazby. Analýza určí, které zboží se prodává společně, a podle toho se potom produkty v obchodě rozmístí. Algoritmus najde všechny podstatné asociace mezi hodnotami v databázi, analytik však musí rozhodnout o jejich relevanci, k čemuž mu dopomáhají statistické testy (Spearmanův, Kendallův apod.). Ukázku použití pravidel, jejichž výstup je ve 2D grafu, zobrazuje obrázek 1.

Použitá data zachycují bonitu klientů. Není žádné překvapení, že pokud klient nemá žádné další úvěry, je jeho credit rating hodnocen kladně, což ukazuje nejsilnější vazba v grafu. Také můžeme vidět, že bydlí-li klient, který si půjčil například na nové auto, v bytě, který si pronajal, další úvěr obvykle nemá. Jistá vazba je také v souvislosti s pohlavím a dalšími úvěry, slabší vztah pak můžeme vypozorovat mezi dalšími úvěry a povoláním. Míra podrobnosti závisí na konfiguraci citlivosti sdružených pravidel, kde se, jednoduše řečeno, nastavuje určitá minimální míra závislosti mezi proměnnými.

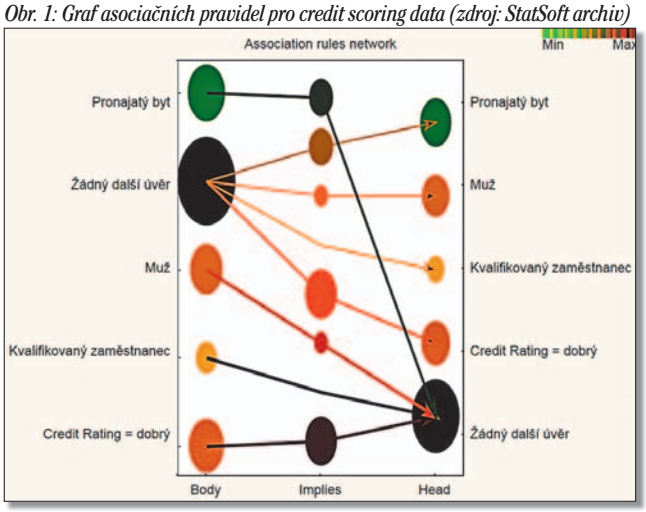
Co je to neuronová síť

Neuronová síť je založena na analogii s lidským mozkem, je tedy tvořena množstvím vzájemně propojených prvků neboli neuronů. Neuron je chápán jako element, který přijímá podněty od ostatních neuronů, jež jsou k němu na vstupu připojeny. Vnitřní algoritmy neuronových síti jsou velice složité. Nás nemusí zajímat až tak to, co se uvnitř sítě děje, jako to, k čemu ji můžeme využít. Jaká je tedy výhoda neuronové sítě?

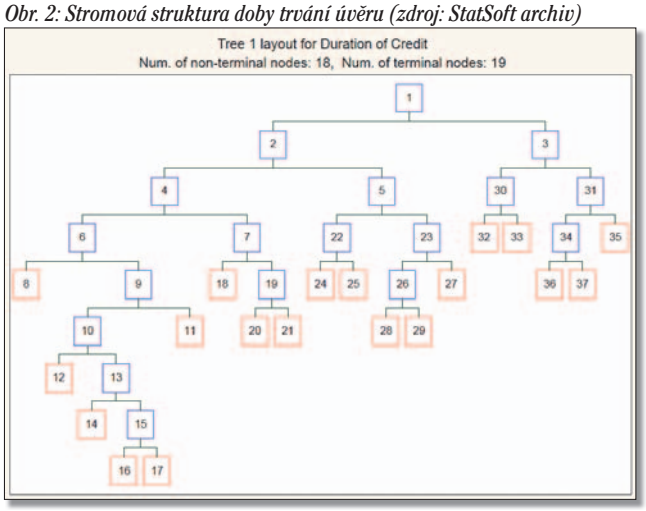
Hlavní předností sítě je učení se z příkladů, které jí dáme k dispozici. Použití sítě v praxi vypadá přibližně následovně: máme k dispozici data, jež chceme klasifikovat, časovou řadu, kterou chceme modelovat, apod. V první fázi si určíme poměr dat tréninkových, testovacích a validačních. Síť načte tréninková data a začne se učit. Doba zpracování algoritmu učení je samozřejmě závislá na objemu datového souboru. Po zpracování první části dat (training sample) se síť pustí do testovací části (testing sample) a bude si ověřovat to, co se naučila. Poslední část dat, sloužící k validaci, je následně použita pro vlastní analýzu.

V praxi není dobré dávat příliš velký objem dat do první fáze. Ano, vzorek zajisté nemůže být příliš malý, ale při velkém testovacím vzorku hrozí tzv. přeučení sítě. To lze vysvětlit na zjednodušeném příkladu: pokud se budete na zkoušku učit málo, test neuděláte, nebudete mít dostatek znalostí, abyste si poradili se všemi příklady, a výsledky

Obr. 1: Graf asociačních pravidel pro credit scoring data (zdroj: StatSoft archiv)



Obr. 2: Stromová struktura doby trvání úvěru (zdroj: StatSoft archiv)



budou nepřesné. Pokud se však budete učit až příliš dlouho, tak ve vašem ostrém testu budete hledat složitosti, které tam nejsou, ztratíte nadhled a při řešení budete dělat chyby. Obvykle proto volíme objem dat v poměru 50/25/25 pro trénování/testování/validaci. Oblíbené typy neuronových sítí pro klasifikaci jsou perceptronová neuronová síť a síť RBF (radial basis function).

Řešení pomocí stromů

Jiným přístupem k otázce klasifikace jsou rozhodovací stromy. Pro vysvětlení principu stromů nám postačí binární stromy, kde jsou data v každém nekoncovém uzlu dále dělena pouze do dvou skupin. Klasifikační strom je budován na základě výukového vzorku dat. Cílem je rozdělit jednotlivé objekty do skupin se stejnou klasifikací pomocí několika jednoduchých pravidel, která vycházejí ze vztahů měřených veličin ke klasifikaci. Strom má několik úrovní, každá úroveň obsahuje buď tzv. uzly, jež se dále rozkládají na základě dělících pravidel, nebo jde o tzv. listy, kde k dělení nedochází a kterým je již přiřazena konkrétní klasifikační třída. Stromovou strukturu ukazuje obrázek 2.

V každém uzlu se stanoví proměnná, pomocí níž nejlépe rozdělíme objekty tohoto uzlu do dvou skupin, jejichž prvky mají v rámci konkrétní skupiny podobné klasifikace, ale tyto klasifikace se liší od klasifikací prvků skupiny druhé. Uchazeče o úvěr můžeme v prvním kroku rozdělit například podle výše příjmu – lidé s příjmem nad osm tisíc korun a lidé s příjmem nižším. Každou z takto vzniklých skupin dělíme dále podle vhodných kritérií. Pro případ, že ne u všech objektů máme k dispozici hodnoty veličiny, podle níž se řídí dělení v některém uzlu, stanoví se pro uzly jedna až tři zástupné proměnné, které využíváme pro dělení v případě chybějících hodnot. Klasifikace nových případů odpovídá většinové klasifikaci vzorových objektů v příslušném listu.

Jak roste strom?

Jako míra kvality stromového modelu se používá některé z vhodných kritérií, jež odráží celkové procento nesprávně klasifikovaných případů. Penalizace za nesprávnou klasifikaci přitom může být odlišná pro každý případ špatné klasifikace (klient, který je nevhodný, ale model ho zahrne mezi vhodné uchazeče, způsobí větší ztrátu než dobrý klient, kterému se rozhodneme neposkytnout úvěr – ohodnocení tedy nejsou symetrická). Analogicky k neuronovým sítím i u stromů je žádoucí, aby nebyly příliš velké nebo příliš

malé. V praxi se používají dva postupy pro získání rozumně obsáhlé stromové struktury: vytvoříme na základě výukových dat úplný binární klasifikační strom, kde v každém jeho listě budou pouze objekty s toutéž klasifikací. Takové schéma je však zbytečně detailní a těžko se interpretuje. Proto hned přistoupíme k ořezání – přecházíme postupně jednotlivé nekoncové uzly a zvažujeme jejich nahrazení listem. Pokud je uzel nahrazen listem, znamená to, že jeho objekty se již dále nedělí a všem novým objektům, které skončí v tomto listu, je přiřazena stejná klasifikace. Druhou možností je přímo vytvoření redukovaného stromu, kdy proces výstavby stromové struktury ukončíme, když případné přidání nových uzlů výrazně nezlepší model.

Pokročilejší dataminingové modely

Kromě již zmíněných metod se můžeme setkat také s modely, které jsou zobecněním nebo kombinací uvedených. Náhodné lesy (boosted trees) představují modely složené z několika (ne nutně binárních) stromů. Každý jednoduchý strom je pouze slabým klasifikátorem, ale jejich kombinace již představuje velmi dobrý model. Výstupy rozdělení stromových modelů mohou sloužit jako vstupy neuronové sítě, dostáváme se tak k termínu meta-learning, kdy vytváříme model neuronové sítě s použitím jiného modelu, jenž předem zpracovává stejná data.

Závěr

Podotkněme, že z uvedených modelů mají pouze stromy snadnou interpretaci – vidíme zde, která kritéria jsou pro klasifikaci rozhodující. Ostatní modely sice mohou často přinášet lepší výsledky než modely tradiční, ale fungují spíše jako černá skříňka. Jejich výklad je třeba hledat (což samo o sobě představuje další statistické úlohy). Bez věcné interpretace našeho modelu totiž nemáme dostatečně silné argumenty, kterými bychom přesvědčili vedení firmy nebo instituce o vhodnosti implementace modelů vytvořených pomocí těchto moderních metod. Přesto jejich oblíbenost stále roste. S rozvojem databází vzrůstá počet uložených dat, který je pro tyto metody rozhodující. Velké objemy a značné množství ukládaných proměnných o každém z nás poskytne dataminingovému nástroji vhodný „výukový“ materiál, na jehož základě jsou potom vytvářeny stále lepší modely. Ty pak odhalují často nečekané vzory chování. ■

Autoři působí ve společnosti StatSoft CR.