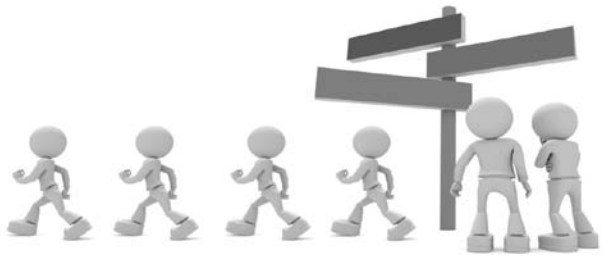
Moderné data miningové metódy

Lenka Blažková

V našom poslednom príspevku v tomto roku sa budeme zaoberať metódami, ktoré svojou povahou nespadajú do štatistiky, a síce postupmi, ktoré sa súhrnne označujú ako prostriedky pre data mining – získavanie znalostí z dát. Pretože tato skupina metód je prinajmenšom rozsiahla, naša užšia téma budú metódy užívané pri riešeniu tzv. **klasifikačných úloh**. Ide o problémy, kedy sa snažíme nájsť pravidla pre rozdelenie objektov do vopred daných skupín. Štandardné štatistické postupy využívané pre takýto typ úloh sú diskriminačná analýza a logistická regresia. Ich modernou alternatívou môžu byť napríklad **klasifikačné rozhodovacie stromy** alebo **neurónové siete**.

O čo ide v klasifikačných úlohách?

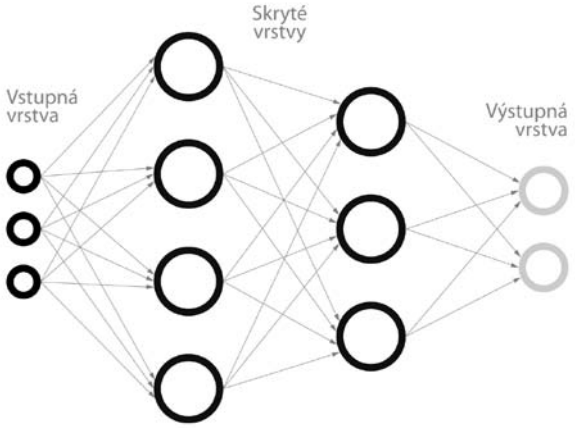
Štruktúra dát, ktoré sú predmetom analýzy, je akási matica alebo tabuľka, ktorej riadky predstavujú jednotlivé objekty (napríklad jednotlivých zákazníkov) a stĺpce odpovedajú meraným vlastnostiam objektov (vek, výška príjmu, rodinný stav, pohlavie, mesačný obrat na účte, ...). Pre každý objekt máme k dispozícii aj jeho klasifikáciu – príslušnosť k jednej z existujúcich skupín (pre účely credit skóringu to sú najčastejšie dobrí a zlí klienti vo vzťahu ku splácaniu úverov, prípadne môže existovať i neutrálna skupi-



na, u ktorej si pre konečné rozhodnutie pýtame ďalšie informácie). Takáto klasifikácia je však dostupná iba pre niektoré objekty. Klasifikačná úloha spočíva v tom, že je treba správne klasifikovať aj objekty nové. Objekty, pre ktoré máme informáciu o príslušnosti k jednej z možných skupín, používame ako učiace (vzorové) dáta. Hľadáme v nich súvislosť medzi klasifikáciou a ostatnými meranými veličinami – na základe analýzy vplyvu ostatných premenných na klasifikáciu objektov vytvoríme model, ktorý aplikujeme na nové dáta. V dnešnej dobe sa k tomuto účelu z moderných data miningových metód používajú neurónové siete a algoritmy založené na rozhodovacích stromoch. Teraz sa budeme bližšie venovať týmto dvom technikám.

Princíp neurónových sietí

Neurónové siete sa snažia napodobňovať činnosť neurónov v ľudskom mozgu. Ich základom je model umelého neurónu, ktorý na základe vstupov (buď z vonkajšku alebo od iných neurónov) generuje výstup. Podľa toho, kde sa neurón nachádza, rozlišujeme **vstupnú**, **skrytú** a **výstupnú** vrstvu neurónov.



Každý zo vstupov má inak veľký vplyv na výstup neurónu – toho sa dosiahne v tzv. **učiacej fáze** algoritmu, kedy sú generované výstupy porovnávané s výstupmi cieľovými (známou klasifikáciou) a kedy sa váhy jednotlivých vstupov na základe učiacich dát postupne nastavujú tak, aby sme dosiahli čo najlepšie zhody – minimalizujú sa odchýlky generovaných a cieľových výstupov. Umelý neurón vyhodnocuje vstupy a jeho výstupom je **aktivačná funkcia**, ktorá je funkciou váženého súčtu vstupov. Hodnota aktivačnej funkcie slúži ako vstup neurónom ďalšej

vrstvy alebo, ak ide o neurón vo výstupnej vrstve, už udáva hľadanú klasifikáciu predpísanú neurónovou sieťou.

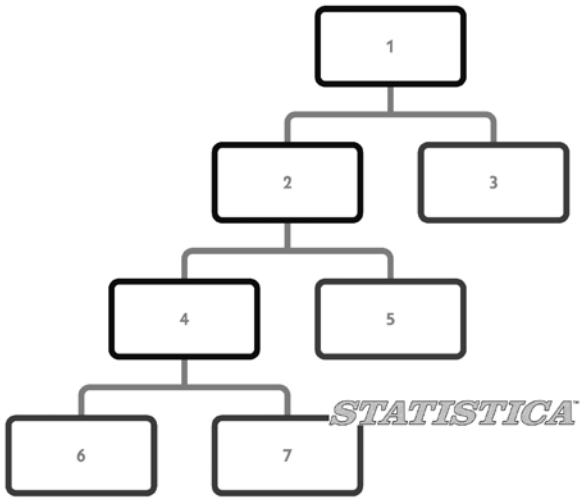
Ideálna neurónová sieť

Výsledná neurónová sieť by nemala byť príliš rozsiahla, pretože potom dochádza k **preučeniu**, kedy sieť nie je schopná zovšeobecnenia pre iné než učiace dáta, ale ani príliš malá, lebo potom totiž nie je schopná dobre rozlišovať medzi objektmi rôznych klasifikačných tried. Pre nájdenie počtu skrytých neurónov a takisto pre úpravu váh v procese učenia existuje mnoho algoritmov, často sa pre nastavenie váh používa algoritmus zvaný **backpropagation**. Obľúbené typy neurónových sietí pre klasifikáciu sú **perceptronová neurónová sieť** a sieť **RBF** (radial basis function).

Riešenia pomocou stromov

Iným prístupom k otázke klasifikácie sú rozhodovacie stromy. Pre vysvetlenie princípu stromov nám postačia **binárne stromy**, kedy sú dáta v každom nekoncovom uzle ďalej delené iba do dvoch skupín.

Klasifikačný strom je budovaný na základe učiacej vzorky dát. Cieľom je jednotlivé objekty rozdeliť do skupín s rovnakou klasifikáciou pomocou niekoľkých jednoduchých pravidiel, ktoré vychádzajú zo vzťahov meraných veličín ku klasifikácii. Strom má niekoľko úrovní, každá úroveň obsahuje tzv. **uzly**, ktoré sa buď ďalej delia na základe deliacich pravidiel, alebo ide o tzv. **listy**, kde k deleniu nedochádza



a ktorým už je priradená konkrétna klasifikačná trieda. Stromovú štruktúru ukazuje nasledujúci obrázok.

V každom uzle sa stanoví premenná, pomocou ktorej vieme najlepšie rozdeliť objekty tohto uzlu do dvoch skupín, ktorých prvky majú v rámci konkrétnej skupiny podobné klasifikácie, ale tieto klasifikácie sa líšia od klasifikácií prvkov skupiny druhej. Uchádzača o úver môžeme v prvom kroku rozdeliť napríklad podľa výšky príjmu – ľudia s príjmom nad 1500 EUR a ľudia s príjmom nižším. Každú z takto vzniknutých skupín delíme ďalej podľa vhodných kritérií. Pre prípad, že nie u všetkých objektov máme k dispozícii hodnoty veličiny, podľa ktorej sa riadi delenie v niektorom uzle, stanovujú sa pre uzly aj jedna až tri **zástupné premenné**, ktoré využívame pre delenie v prípade chýbajúcich hodnôt. Klasifikácia nových prípadov odpovedá väčšinovej klasifikácii vzorových objektov v príslušnom liste.

Ako rastie strom?

Ako miera kvality stromového modelu sa používa niektoré vhodné kritérium, ktoré odráža celkové percento nesprávne klasifikovaných prípadov. Penalizácia za nesprávnu klasifikáciu pritom môže byť odlišná pre každý prípad zlej klasifikácie (klient, ktorý je zlý, ale model ho zahrnie medzi dobrých uchádzačov, spôsobí väčšiu stratu ako dobrý klient, ktorému sa rozhodneme neposkytnúť úver – ohodnotenie teda nie je symetrické).

Analogicky k neurónovým sieťam, i u stromov je žiaduce aby neboli priveľké alebo príliš malé. V praxi sa používajú dva postupy pre získanie rozumne veľkej stromovej štruktúry: Vytvoríme na základe učiacich dát úplný binárny klasifikačný strom, kedy v každom jeho liste budú iba objekty s tou istou klasifikáciou. Takáto schéma je však zbytočne detailná a ťažko sa interpretuje. Preto hneď pristúpime k orezávaniu – prechádzame postupne jednotlivé nekoncové uzly a zvažujeme ich nahradenie listom. Ak je uzol nahradený listom, znamená to, že jeho objekty sa už ďalej nedelia, a všetkým novým objektom, ktoré skočia v tomto liste, je priradená rovnaká klasifikácia. Druhou možnosťou je priamo vytvorenie redukovaného stromu, kedy proces výstavby stromovej štruktúry ukončíme vtedy, keď prípadné pridanie nových uzlov výrazne nezlepší model.

Pokročilejšie data miningové modely

Okrem už spomenutých neurónových sietí a binárnych stromov sa môžete stretnúť aj s modelmi, ktoré sú zovšeobecnením alebo kombináciou týchto. **Náhodné lesy** alebo **boosted trees** predstavujú

modely zložené z niekoľkých (nie nutne binárných) stromov. Každý jednoduchý strom je iba slabým klasifikátorom, ale kombinácia týchto jednoduchých stromov predstavuje už veľmi dobrý model.

Výstupy klasifikačných stromových modelov môžu slúžiť ako vstupy neurónovej siete, dostávame sa tak k termínu **meta-learning**, kedy vytvárame model neurónovej siete s použitím iného modelu, ktorý vopred spracováva tie isté dáta.

Záver

Podotkneme, že z uvedených modelov majú iba stromy ľahkú interpretáciu – vidíme, ktoré kritériá sú pre klasifikáciu rozhodujúce. Ostatné modely síce môžu prinášať často lepšie výsledky ako modely tradičné, ale fungujú skôr ako čierna skrinka. Ich interpretáciu je však treba hľadať (čo samo o sebe predstavuje ďalšiu štatistickú úlohu). Bez vecnej interpretácie nášho modelu totiž nemáme dostatočne silné argumenty, ktorými by sme presvedčili vedenie firmy alebo inštitúcie o vhodnosti implementácie modelov vytvorených pomocou týchto moderných metód.

Autorkou článku je Mgr. Lenka Blažková, odborná konzultantka firmy StatSoft CR