SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2017
ročník/volume 27

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

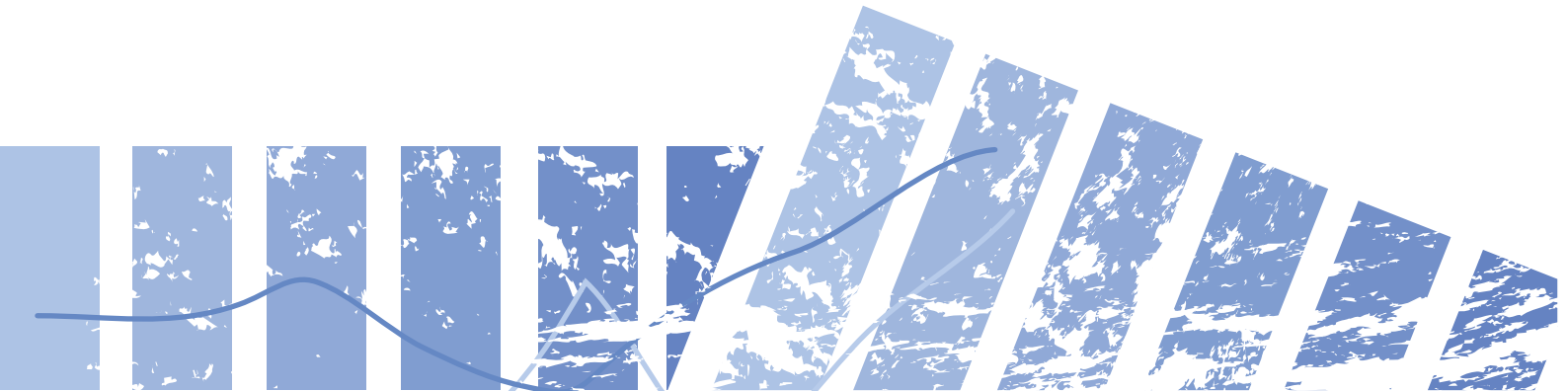
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 6

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 60 – 76

Dátum vydania/Publication date: 15. júl 2017/July 15, 2017



Viera LABUDOVÁ
Katedra štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave

ROZHODOVACIE STROMY AKO PREDIKTÍVNA MODELOVACIA TECHNIKA

DECISION TREES AS A PREDICTIVE MODELING METHOD

ABSTRAKT
Rozhodovacie stromy sú silným nástrojom, ktorý sa používa na predikciu a klasifikáciu. Príťažlivosť metód založených na rozhodovacích stromoch je daná skutočnosťou, že rozhodovacie stromy predstavujú pravidlá. Ak sa rozhodovací strom používa na klasifikáciu (cieľová premenná je kategoriálna), nazýva sa klasifikačný strom. Ak sa použije na regresné úlohy (cieľová premenná je spojitá), nazýva sa regresný strom. Článok sa venuje opisu štruktúry rozhodovacích stromov a základným algoritmom na konštrukciu rozhodovacích stromov.

ABSTRACT
Decision trees are powerful tools for classification and prediction. The attractiveness of tree-based methods is largely due to the fact that decision trees represent rules. When a decision tree is used for classification tasks (the target variable is categorical), it is referred to as a classification tree. When it is used for regression tasks (the target variable is continuous), it is called a regression tree. This article describes the structure of decision trees and the basic algorithm for their construction.

KĽÚČOVÉ SLOVÁ
rozhodovacie stromy, entropia, Giniho index, algoritmy rozhodovacích stromov

KEY WORDS
decision trees, entropy, Gini index, decision tree algorithms

1. ÚVOD

Modely rozhodovacích stromov založené na niekoľkých vstupných a jednej výstupnej premennej patria do skupiny viacrozmerných štatistických metód. Na základe hodnôt vstupných (vysvetľujúcich) premenných sa odhaduje hodnota spojitej závislej (výstupnej) premennej alebo sa jednotlivé objekty zaraďujú do príslušných skupín zodpovedajúcich kategóriám závislej premennej. Rozhodovacie stromy možno preto považovať za alternatívny prístup k lineárnej regresnej analýze, ak je závislá premenná číselná spojitá, alebo k logistickej regresii a diskriminačnej analýze, ak je závislá premenná kategoriálna [14]. Rozhodovacie stromy možno zaradiť tiež do skupiny hierarchických zhlukovacích metód, keďže ich výstupom sú disjunktné podmnožiny pôvodného súboru objektov [5].

Rozhodovacie stromy sa používajú v situáciách, keď potrebujeme predikovať hodnoty spojitej závislej premennej, alebo v prípadoch, keď predikujeme príslušnosť objektov do vopred zvolených tried.

V literatúre sa môžeme stretnúť s rôznymi definíciami rozhodovacieho stromu.

Rozhodovací strom je štruktúra, ktorá sa využíva na rozdelenie veľkého súboru prípadov v databáze na menšie súbory prípadov pri postupnej aplikácii jednoduchých rozhodovacích pravidiel. Rozhodovací strom pozostáva zo súboru pravidiel (predpisov)[1] na rozdelenie veľkej heterogénnej populácie do menších, homogénnejších skupín s rešpektovaním príslušnej výstupnej premennej [2].

Rozhodovací strom predstavuje reprezentáciu rozhodovacej procedúry na klasifikáciu prípadov do príslušných tried. Je to grafová štruktúra vo forme stromu obsahujúca koreňový uzol, nelistové a listové uzly. Uzly reprezentujú triedu alebo testovací znak. Hrany reprezentujú hodnoty testovacieho znaku [8].

2. GENEROVANIE ROZHODOVACIEHO STROMU

Rozhodovacie stromy sa generujú postupom, ktorý sa nazýva TDIDT (*top down induction of decision trees*) – indukcia rozhodovacích stromov zhora nadol. Algoritmus na generovanie rozhodovacieho stromu sa začína na trénovacej množine, ktorá sa nazýva aj priestor prípadov, množina prípadov alebo základný priestor. Ten je tvorený hodnotami vstupných premenných (znakov) X_1, X_2... X_k, ktoré môžu byť číselné (diskrétne, spojité) aj slovné, a hodnotami výstupnej premennej Y, zistených na množine prípadov (objektov). Hodnoty výstupnej premennej vytvárajú triedy.

Základný priestor sa v procese generovania rozhodovacieho stromu delí na podpriestory, ktoré sú charakterizované hodnotami testovacích znakov. Delenie sa uskutočňuje rekurzívne, kým nie je splnená tzv. ukončovacia podmienka. Pri tomto postupe sa množina prípadov postupne delí na menšie a menšie podmnožiny (podpriestory), v ktorých prevládajú prípady jednej triedy alebo prípady s podobnou hodnotou znaku.

Tento algoritmus, ktorý sa považuje za všeobecný postup generovania rozhodovacieho stromu zhora nadol, môžeme zapísať takto [8]:

1. Ak je pre každý podpriestor splnené ukončovacie kritérium, generovanie sa ukončí.
2. Inak:
3. Zvolí sa podpriestor obsahujúci prípady klasifikované do viacerých tried.
4. Pre zvolený podpriestor sa vyberie jeden testovací znak, ešte nepoužitý pre daný podpriestor prípadov.
5. Zvolený podpriestor prípadov sa rozdelí na ďalšie podpriestory podľa hodnôt zvoleného testovacieho znaku.

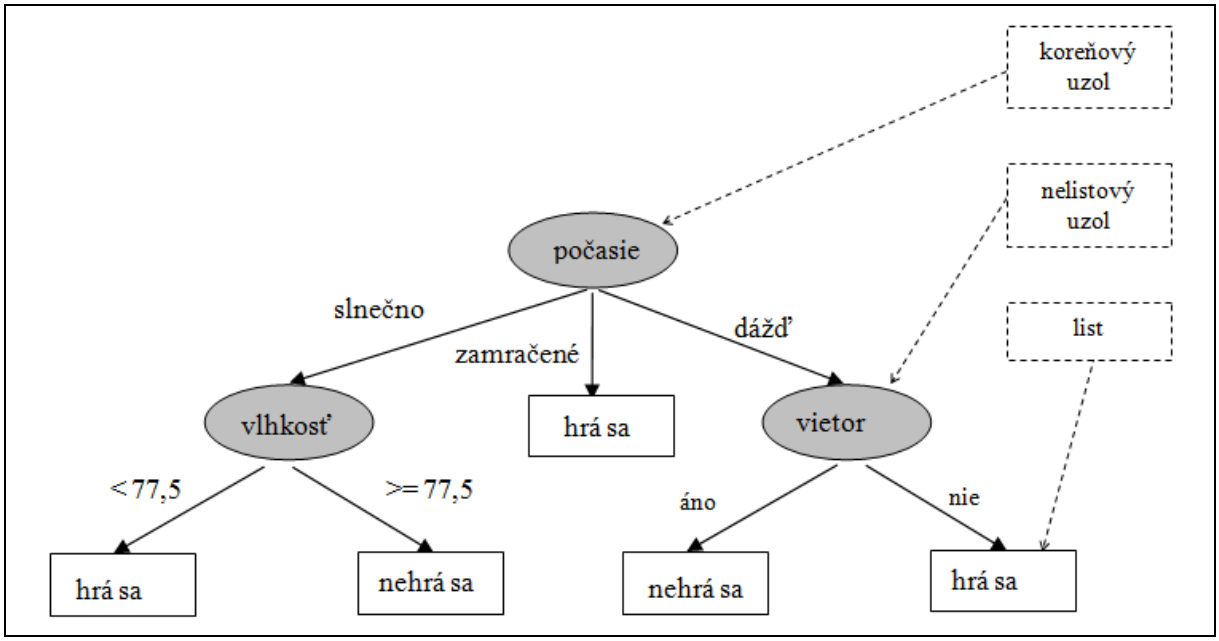
Krok 1 zodpovedá prípadu, keď generovanie rozhodovacieho stromu je už ukončené. Krok 2 pokračuje v generovaní, keď expanduje jeden listový uzol o jednu úroveň.

Uvedený algoritmus okrem pravidla budovania stromu zhora nadol zahŕňa aj pravidlo „rozdeľuj a panuj“, ktoré sa uplatňuje pri indukovaní podstromov, pri ktorom sa úlohy delia na podúlohy. Pri tomto postupe sa množina prípadov postupne delí na menšie a ešte menšie podmnožiny (podpriestory), v ktorých prevládajú prípady jednej triedy.

[1] *Pravidlami sú v tomto prípade vzťahy medzi premennými.*

Rozhodovacie stromy sa zobrazujú najčastejšie v grafickej podobe, pričom existuje veľa možností ich znázornenia. Obrázok 1 ilustruje jednu z možností zobrazenia rozhodovacieho (klasifikačného) stromu.

Obrázok č. 1: Grafické zobrazenie rozhodovacieho stromu



Zdroj: [10], vlastné spracovanie

Klasifikačný strom na obrázku 1 vznikol na trénovacej množine uvedenej v tabuľke 1. Prípadmi sú dni, počas ktorých sa sledovali atribúty počasia (vysvetľujúce premenné), závislou, modelovanou premennou je premenná hra, ktorá nadobúda dve obmeny: hrá sa, nehrá sa.

Tabuľka č. 1: Trénovacia množina prípadov

Vstupné premenné				Cieľová premenná
počasie	**teplota**	**vlhkosť**	**vietor**	**hra**
Slnečno	29	85	nie	nehrá sa
Slnečno	27	90	áno	nehrá sa
Zamračené	28	78	nie	hrá sa
Dážď	21	96	nie	hrá sa
Dážď	20	80	nie	hrá sa
Dážď	18	70	áno	nehrá sa
Zamračené	18	65	áno	hrá sa
Slnečno	22	95	nie	nehrá sa
Slnečno	21	70	nie	hrá sa
Dážď	24	80	nie	hrá sa
Slnečno	24	70	áno	hrá sa
Zamračené	22	90	áno	hrá sa
Zamračené	27	75	nie	hrá sa
Dážď	22	80	áno	nehrá sa

Zdroj: [10]

Vytvorený klasifikačný strom je trojúrovňovou[2] hierarchickou štruktúrou, ktorá má osem uzlov. V koreňovom uzle sa nachádza celá množina prípadov (objekty opísané hodnotami vstupných premenných a hodnotami výstupnej premennej). Nelistové (neterminálové, medziľahlé uzly) sú v tomto grafe zobrazené oválom. Tie reprezentujú testovací znak (vetviaci znak). Zhora nadol orientované hrany (vetvy stromu), ktoré vychádzajú z týchto uzlov, zodpovedajú kategóriám testovacích znakov. Podpriestory, ktoré sa už ďalej nedelia, sa nazývajú listy (listové, terminálové uzly), tie sú zobrazené obdĺžnikom. Listy najčastejšie obsahujú informáciu o zaradení objektu do klasifikačnej triedy (v prípade klasifikačných stromov) alebo informáciu o odhadnutej strednej hodnote modelovanej premennej (v prípade regresných stromov). Okrem toho obsahujú aj informáciu o početnostiach príslušných tried závislej premennej.

Keďže sa každý trénovací príklad asociuje s jedným listovým uzlom, stromovú štruktúru možno prepísať do súboru rozhodovacích (klasifikačných, produkčných) pravidiel. Každé klasifikačné pravidlo obsahuje opis jednej cesty od koreňového uzla po niektorý listový uzol. Pravá strana pravidla obsahuje názov triedy zodpovedajúcej listovému uzlu, v ktorom sa cesta končí. Do tejto triedy je zaradený každý trénovací príklad spĺňajúci podmienky ľavej strany pravidla. Ako príklad rozhodovacích pravidiel uvedieme pravidlá, do ktorých je prepísaná stromová štruktúra uvedeného rozhodovacieho stromu (obr. 2).

Obrázok č. 2: Rozhodovacie pravidlá

```
Node = 4
if pocasie IS ONE OF: ZAMRAČENÉ  then Tree Node Identifier = 4
Number of Observations = 4  Predicted: hra = nehrá sa  = 0.00  Predicted: hra = hrá sa  = 1.00

Node = 5
if vlhkost < 77.5 AND pocasie IS ONE OF: SLNEČNO or MISSING  then Tree Node Identifier = 5
Number of Observations = 2  Predicted: hra = nehrá sa  = 0.00  Predicted: hra = hrá sa  = 1.00

Node = 6
if vlhkost >= 77.5 or MISSING AND pocasie IS ONE OF: SLNEČNO or MISSING then Tree Node
Identifier  = 6
Number of Observations = 3  Predicted: hra = nehrá sa  = 1.00  Predicted: hra = hrá sa  = 0.0

Node = 7
if vietor IS ONE OF: ÁNO AND pocasie IS ONE OF: DÁŽĎ then  Tree Node Identifier = 7
Number of Observations = 2  Predicted: hra = nehrá sa  = 1.00  Predicted: hra = hrá sa  = 0.00

Node = 8
if vietor IS ONE OF: NIE or MISSING AND pocasie IS ONE OF: DÁŽĎ then  Tree Node
Identifier = 8
Number of Observations = 3  Predicted: hra = nehrá sa  = 0.00  Predicted: hra = hrá sa  = 1.00
```

Zdroj: vlastné spracovanie, SAS EM

Ak je výstupná premenná kategoriálna, každý listový uzol predstavuje niektorú z kategórií, tried výstupnej premennej. Vtedy hovoríme o klasifikačných stromoch. Ak je výstupná premenná spojitá, každý list reprezentuje odhadnutú hodnotu výstupnej premennej. V takomto prípade hovoríme o regresných stromoch. V uvedenom

[2] *Úroveň obsahujúca koreňový uzol sa považuje za nultú.*

príklade má modelovaná premenná dve triedy, preto listy reprezentujú jednu z tried znaku hra: hrá sa, nehrá sa.

2.1. Kritériá výberu testovacích znakov

Kritériá výberu testovacích znakov pre klasifikačné stromy

Kritérium na výber premennej, ktorá sa použije na príslušnej úrovni vetvenia, závisí od charakteru výstupnej premennej. Základná idea rastu stromu súvisí s teóriou čistoty údajov. Kritériom výberu vetvenia je zvyšovanie čistoty dcérskych uzlov[3]. Výber testovacieho znaku sa môže uskutočniť rôznymi postupmi. Ak je výstupná premenná kategoriálna, používa sa pri výbere testovacieho znaku Giniho index, entropia, informačný zisk alebo chí-kvadrát test nezávislosti. Ak je výstupná premenná spojitá, jednou z možností je kategorizácia jej hodnôt a použitie niektorej z už spomenutých mier. Pri zachovaní jej pôvodného charakteru sa uplatňuje redukcia rozptylu alebo F-test.

Pri posudzovaní kvality delenia (vetvenia) sa využívajú miery čistoty vzniknutých dcérskych uzlov, ktoré vychádzajú z entropie: informačný zisk (*Information Gain*) a pomerný informačný zisk.

Entropia

Uvažujme o trénovacej množine *n* prípadov. Každý prípad je opísaný hodnotou vstupného znaku *A*[4] a hodnotou výstupného znaku *Y*. Nech nadobúda vstupný znak hodnoty a_i ($i = 1,2,\ldots k$) a nech má výstupný znak *m* rôznych hodnôt – tried y_j, ($j = 1,2,\ldots m$). Pravdepodobnosť výskytu triedy y_j, ($j = 1,2,\ldots m$) výstupného znaku *Y* označme p_j ($j = 1,2,\ldots m$).

Entropiu výstupného znaku *Y* vyjadríme takto:

$$H(Y) = -\sum_{j=1}^{m} \left(p_j \log_2 p_j \right), \tag{1}$$

kde p_j je pravdepodobnosť výskytu *j*-tej triedy výstupného znaku *Y*.

Pravdepodobnosť p_j môžeme odhadnúť pomocou relatívnej početnosti $\frac{n_j}{n}$, kde n_j je absolútna početnosť triedy y_j, $j = 1,2,\ldots m$ v množine trénovacích prípadov. Vzťah (1) potom upravíme na tvar

$$H(Y) = -\sum_{j=1}^{m} \left(\frac{n_j}{n} \log_2 \frac{n_j}{n} \right). \tag{2}$$

[3] *Za čistý uzol sa považuje taký, ktorý obsahuje len prípady jednej triedy výstupného znaku.*

[4] *A budeme považovať za kategoriálnu premennú. Vstupnými premennými môžu byť aj spojité premenné. Pri tvorbe rozhodovacieho stromu nie je možné vytvárať vetvy pre každú hodnotu premennej, preto dochádza v procese rastu stromu ku kategorizácii hodnôt spojitých premenných.*

Ak má výstupný znak *Y* len dve kategórie, entropia nadobúda minimálnu hodnotu 0 vtedy, ak všetky prípady patria do tej istej triedy. Ak je početnosť obidvoch tried výstupného znaku rovnaká, entropia dosahuje maximálnu hodnotu 1.

Použitím vstupného znaku *A* rozdelíme množinu prípadov do *k* tried (*i*-tá trieda obsahuje všetky prípady s hodnotou a_i, $i = 1, 2, \ldots k$). Očakávaná entropia znaku *A* je vyjadrená vzťahom

$$H(A) = \sum_{i=1}^{k} p_i H(a_i) = \sum_{i=1}^{k} \frac{n(a_i)}{n} H(a_i), \tag{3}$$

kde $n(a_i)$ je počet prípadov trénovacej množiny, ktoré nadobúdajú hodnotu a_i znaku *A*, $H(a_i)$ je entropia na množine prípadov, ktoré majú hodnotu a_i znaku *A,* a *n* je počet všetkých prípadov tejto množiny. $H(a_i)$ vypočítame takto:

$$H(a_i) = -\sum_{j=1}^{m} \frac{n_j(a_i)}{n(a_i)} log_2 \frac{n_j(a_i)}{n(a_i)}, \tag{4}$$

kde $n_j(a_i)$ je počet prípadov množiny *j*-tej triedy výstupného znaku *Y*, ktoré majú hodnotu a_i znaku *A*. Na vetvenie množiny sa vyberá vstupný znak, pre ktorý je hodnota očakávanej entropie najmenšia.

Informačný zisk

Pre entropiu, ktorá je mierou nečistoty, je stanovený informačný zisk $Z(A)$ znaku *A* takto:

$$Z(A) = H(Y) - H(A). \tag{5}$$

Informačný zisk znaku *A* je očakávané zmenšenie entropie zapríčinené rozdelením prípadov na základe kategórií znaku *A*. Informačný zisk na množinách vytvorených vetvením na základe kategórií premennej *A* je definovaný ako rozdiel entropie vyčíslenej na celej množine údajov $H(Y)$ a entropie $H(A)$ na podmnožinách, ktoré dostaneme vetvením uzla (množiny prípadov, ktoré uzol obsahuje) na základe kategórií premennej A. Na vetvenie sa vyberie znak s najvyššou hodnotou informačného zisku.

Pomerný informačný zisk

Pomerný informačný zisk na rozdiel od entropie a informačného zisku zohľadňuje počet hodnôt znaku *A*, ktorý sa použil pri vetvení. Pomerný informačný zisk $PZ(A)$ je definovaný ako podiel informačného zisku $Z(A)$ a tzv. vetvenia $V(A)$

$$PZ(A) = \frac{Z(A)}{V(A)}, \tag{6}$$

kde je vetvenie $V(A)$ definované takto:

$$V(A) = -\sum_{i=1}^{k} \frac{n(a_i)}{n} log_2 \frac{n(a_i)}{n} \quad . \tag{7}$$

Giniho index

Giniho index má pri výbere premenných a postupnosti ich zaraďovania v procese generovania stromu podobnú funkciu ako entropia. Giniho index je definovaný

$$G = 1 - \sum_{j=1}^{m} p_j^2 \ , \tag{8}$$

kde p_j je pravdepodobnosť výskytu *j*-tej triedy výstupného znaku *Y*.

Podobne ako pri entropii odhadneme pravdepodobnosti pomocou relatívnych početností a vzťah na výpočet Giniho indexu na celej množine prípadov potom upravíme na tvar

$$G(Y) = 1 - \sum_{j=1}^{m} \left(\frac{n_j}{n} \right)^2 \quad . \tag{9}$$

Očakávanú hodnotu Giniho indexu pre znak *A* určíme analogicky ako pri entropii

$$G(A) = \sum_{i=1}^{k} \frac{n(a_i)}{n} G(a_i) \ , \tag{10}$$

kde $G(a_i)$ je Giniho index na množine prípadov, ktoré majú hodnotu a_i znaku *A*

$$G(a_i) = 1 - \sum_{j=1}^{m} \left(\frac{n_j(a_i)}{n(a_i)} \right)^2 \quad . \tag{11}$$

Na vetvenie sa použije znak s najmenšou očakávanou hodnotou Giniho indexu.

Tak ako pri entropii sme definovali informačný zisk, aj pri Giniho indexe môžeme zaviesť podobnú mieru, ktorou je redukcia nečistoty

$$Z_G(A) = G(Y) - G(A) \, . \tag{12}$$

Pri vetvení sa vyberie znak, ktorého použitie pri vetvení vedie k najväčšej redukcii nečistoty v uzle.

Alternatívne možno na vetvenie stromu použiť aj chí-kvadrát test nezávislosti. Na vetvenie na príslušnej úrovni vetvenia sa použije znak, ktorý má najväčšiu asociáciu

s výstupným znakom. Sila asociácie sa porovnáva pomocou *p*-hodnoty testu nezávislosti. Na vetvenie sa vyberie znak, pre ktorý je *p*-hodnota najmenšia.

Kritériá výberu testovacích znakov pre regresné stromy

Iné spôsoby výberu znakov na vetvenie sa používajú pri regresných stromoch, ktoré modelujú spojitú výstupnú premennú. Regresné stromy sa používajú na odhad očakávanej hodnoty výstupného znaku. Listové uzly v regresných stromoch obsahujú priemernú hodnotu výstupného znaku pre prípady v danom uzle. V regresných stromoch sa na voľbu znaku na vetvenie množiny prípadov používa redukcia smerodajnej odchýlky výstupného znaku alebo F-test.

Redukcia smerodajnej odchýlky

V regresných stromoch možno považovať redukciu smerodajnej odchýlky výstupného znaku za alternatívu k informačnému zisku. Smerodajná odchýlka hodnôt výstupného znaku je na množine *n* prípadov vyjadrená vzťahom

$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (i = 1,2,\ldots n)\,. \tag{13}$$

Použitím vstupného znaku *A* rozdelíme množinu *n* prípadov do *k* tried (*i*-tá trieda obsahuje všetky prípady s hodnotou a_i, ($i = 1,2,\ldots k$). Očakávaná smerodajná odchýlka na podmnožinách, ktoré vzniknú vetvením množiny prípadov na základe hodnôt a_i znaku *A*, je určená vzťahom

$$s_y(A) = \sum_{i=1}^{k} \frac{n(a_i)}{n} s_y(a_i)\,, \tag{14}$$

kde $n(a_i)$ je počet prípadov množiny, ktoré nadobúdajú hodnotu a_i znaku *A*, $s_y(a_i)$ je smerodajná odchýlka výstupného znaku *Y* na množine prípadov, ktoré majú hodnotu a_i znaku *A*, a *n* je počet prípadov danej množiny.

Namiesto informačného zisku sa na výber znaku používa redukcia smerodajnej odchýlky

$$Zs_y(A) = s_y - s_y(A)\,. \tag{15}$$

Na vetvenie sa vyberie znak, pre ktorý je redukcia smerodajnej odchýlky (15) najväčšia.

F- test

Pri tomto teste sa porovnávajú stredné hodnoty μ_i ($i = 1,2,\ldots k$) výstupného znaku *Y* na podmnožinách, ktoré dostaneme rozdelením množiny prípadov podľa hodnôt vetviaceho znaku *A*. Pri tomto teste sa overuje platnosť nulovej hypotézy

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ pre $k > 2$

oproti alternatívnej hypotéze

H_1 : aspoň dve stredné hodnoty sa nerovnajú.

Hodnotu testovacej štatistiky možno za predpokladu platnosti nulovej hypotézy vypočítať podľa vzťahu

$$F = \frac{\frac{\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n(a_i)}{k-1}}{\frac{\sum_{i=1}^{k} \sum_{j} (y_{ij} - \bar{y}_i)}{n-k}}, \qquad (16)$$

kde $\bar{y}_i$ je priemerná hodnota znaku *Y* na množine prípadov, ktoré majú hodnotu a_i znaku *A*, $\bar{y}$ je priemerná hodnota znaku *Y* na množine všetkých prípadov, $n(a_i)$ je počet prípadov, ktoré nadobúdajú hodnotu a_i znaku A, y_{ij} sú hodnoty znaku *Y* na množine prípadov, ktoré majú hodnotu a_i znaku *A*, *n* je početnosť množiny prípadov (uzla), ktorú vetvíme, a *k* je počet hodnôt znaku A.

Príslušná testovacia štatistika má Fisherovo rozdelenie s počtom stupňov voľnosti $v_1 = k - 1$ a $v_2 = n - k$.

Nulovú hypotézu na úrovni významnosti *α* zamietneme vtedy, keď $F \geq F_{1-\alpha}(v_1, v_2)$, kde $F_{1-\alpha}(v_1, v_2)$ je (1–α)100-percentný kvantil Fisherovho rozdelenia so stupňami voľnosti $v_1 = k - 1$ a $v_2 = n - k$.

Na vetvenie vyberieme znak s najväčšou hodnotou *F* štatistiky, respektíve s najmenšou *p*-hodnotou *F*-testu. Keď na príslušnej úrovni vetvenia pre žiadny znak nezamietneme H_0, vetvenie skončíme.

Podobne možno na vetvenie použiť aj chí-kvadrát test nezávislosti.

3. ALGORITMY GENERUJÚCE ROZHODOVACIE STROMY

V praxi sa využívajú rôzne softvérové balíky, ako napr. Salford Systems CART, IBM SPSS Modeler, Rapid Miner, SAS Enterprise Miner, Matlab, R, Weka atď., v ktorých sú implementované algoritmy na generovanie rozhodovacích stromov.

Algoritmy generujúce rozhodovacie stromy sa líšia aplikovaným kritériom, ktoré sa používa pri výbere vetviacej premennej, druhom vetvenia (binárne, viacnásobné vetvenie), spôsobom narábania s chýbajúcimi hodnotami, charakterom cieľovej premennej a nastaveniami regulujúcimi rast stromu.

Prvým počítačovo implementovaným algoritmom bol algoritmus AID (*Automatic Iteraction Detection*). Vyvinuli ho v roku 1963 John Sonquist a James Morgan [15].

Na modelovanie hodnôt výstupnej premennej využíva binárne vetvenie. Vstupnými premennými môžu byť nominálne aj ordinálne premenné, výstupná premenná je spojitá. Pri generovaní stromu rozdeľuje algoritmus najskôr uzly, v ktorých je súčet štvorcov odchýlok hodnôt výstupnej premennej od priemeru najväčší. Vetvenie sa

ukončí, ak je pokles súčtu štvorcov odchýlok nižší ako hraničná hodnota, ktorá sa rovná súčinu zvolenej konštanty a celkovej sumy štvorcov odchýlok.

Nasledovníkom tohto algoritmu bol klasifikačný algoritmus THAID (*Theta-Automatic Iteraction Detection*), vyvinutý Jamesom N. Morganom a Robertom C. Messengerom v roku 1973 [3].

Jedným z najstarších a súčasne jedným z najrozšírenejších algoritmov v komerčnej oblasti je CHAID (*CHi-squared Automatic Interaction Detection*). Jeho autor V. Gordon Kass [6] zdokonalil predchádzajúce algoritmy AID a THAID.

CHAID sa používa na modelovanie nominálnej výstupnej premennej, pričom využíva nominálne aj ordinálne vysvetľujúce premenné[5]. Na rozdiel od systému AID generuje nebinárny strom, pričom delenie údajov v jednotlivých uzloch je rekurzívne, t. j. každý uzol sa delí podľa rovnaného predpisu. Na delenie pozorovaní a výber vetviacich znakov využíva chí-kvadrát test.

V koreňovom uzle sa pre každý vstupný znak A_i vytvorí kontingenčná tabuľka rozmerov k x l_i, kde k je počet hodnôt (kategórií) výstupnej premennej Y a l_i je počet hodnôt (kategórií) vstupnej premennej A_i. Na podtabuľkách rozmerov k x 2, ktoré sa vytvárajú pre každú kombináciu dvojíc hodnôt vstupnej premennej A_i, sa pomocou chí-kvadrát testu nezávislosti testuje podobnosť týchto dvoch hodnôt, kategórií. Postupne nastáva zhlukovanie tých dvojíc kategórií vstupnej premennej, pre ktoré je výsledok chí-kvadrát testu štatisticky nevýznamný, a to v poradí rastúcej hodnoty chí--kvadrát štatistiky. Po každom zlúčení kategórií sa prepočítava hodnota chí-kvadrát štatistiky vytvorenej tabuľky. Po ukončení zhlukovania sa hľadá najlepšie vetvenie pre kategórie, ktoré vznikli zlúčením aspoň troch pôvodných kategórií vstupnej premennej. Ak je výsledok chí-kvadrát testu štatisticky významný, uskutoční sa dané vetvenie, ak nie je výsledok štatisticky významný, zachová sa táto zlúčená kategória a prejde sa na ďalšiu premennú. Po dokončení optimálneho zlučovania kategórií pre každú vysvetľujúcu premennú sa vyberie najvhodnejšia premenná na vetvenie, a to na základe výsledku chí-kvadrát testu (p-hodnoty testu po Bonferroniho korekcii). Začiatkom 90. rokov minulého storočia vytvoril Barry de Ville algoritmus Exhaustive CHAID, ktorý uskutočňuje podrobnejšie prehľadávanie. Výsledkom je strom s väčším počtom vetiev [3].

Algoritmus ID3 (*Iterative Dichotomizer 3*) [11] je klasickým príkladom algoritmu, ktorý buduje rozhodovací strom metódou zhora nadol TDIDT. Strom vytvorený s využitím algoritmu ID3 pracuje ako klasifikátor, pri ktorom je výstupná premenná kategoriálna. Kategoriálnymi sú aj vstupné premenné. Ako kritérium výberu deliacich znakov využíva entropiu. Na každej úrovni vetvenia sa zo všetkých potenciálnych vstupných premenných vyberie tá, ktorej použitím nastane rozštiepenie množiny (materského uzla) na také podmnožiny (dcérske uzly), na ktorých je celková entropia najmenšia. Vetvenie sa končí, ak každý list obsahuje pozorovania patriace do jednej triedy, t. j. pozorovania nadobúdajú rovnakú hodnotu výstupnej premennej (hodnota entropie v každom liste je nulová). Aby sa dosiahol čo najjednoduchší strom, entropia by mala čo najrýchlejšie klesnúť na nulovú hodnotu. Strom generovaný algoritmom ID3 vedie k maximálnemu poklesu entropie lokálne v každom kroku. Algoritmom

[5] *Vstupné spojité premenné sa kategorizujú, pričom sa vytvárajú približne rovnako početné intervaly.*

dokáže vytvoriť stromy s vysokým stupňom generalizácie. Je použiteľný len pri riešení neinkrementálnych úloh[6].

V roku 1986 vytvorili matematici Schlimmer a Fisher algoritmus ID4, ktorý bol inkrementálnou modifikáciou algoritmu ID3. Oveľa známejším je algoritmus ID5R, ktorý bol odvodený priamo z algoritmu ID4.

Algoritmus ID5R (*Inductive Dichotomizer 5 Recursive*) je inkrementálnou modifikáciou algoritmu ID3, pričom sa pri ňom nevyskytujú problémy algoritmu ID4. Využíva sa v situáciách, keď nie sú známe všetky trénovacie prípady naraz, do trénovacej množiny sa pridávajú postupne. Ak by sa v takejto situácii použil niektorý neinkrementálny algoritmus, viedlo by to po príchode každého nového prípadu k zrušeniu už existujúceho stromu a indukcia stromu by musela prebehnúť od začiatku. Pri novej indukcii by sa nevyužili informácie získané v predchádzajúcich krokoch. Pri inkrementálnej indukcii každé pridanie nových prípadov vedie k modifikácii už vytvoreného rozhodovacieho stromu. Rozhodovací strom sa rekurzívne aktualizuje pod aktuálnym uzlom pozdĺž vetvy zodpovedajúcej tej hodnote znaku, ktorá sa vyskytla v novom trénovacom prípade. Algoritmus ID5R využíva pri výbere deliacich znakov entropiu, resp. informačný zisk.

Algoritmus C4.5 pracuje na podobnom princípe ako ID3 [13]. Ako vstupné premenné používa nominálne aj číselné spojité premenné, dokáže pracovať s pozorovaniami, pri ktorých chýbajú hodnoty niektorých premenných. Ide o neinkrementálny algoritmus, ktorý buduje strom zhora nadol. Aj pri tomto algoritme musí byť na vytvorenie perfektného rozhodovacieho stromu splnená podmienka neprotirečivosti trénovacích prípadov. Na výber testovacej podmienky využíva pomerný informačný zisk. Algoritmus C4.5 pracuje iba v textovom režime. V oblasti strojového učenia sa považuje za štandard tvorby rozhodovacích stromov. Jeho najnovšia verzia je implementovaná ako algoritmus C5.0 a jeho unixový duplikát See 5.

4. UKONČENIE RASTU STROMU, PREREZÁVANIE ROZHODOVACÍCH STROMOV

Rast, vetvenie stromu sa ukončí, ak je splnená niektorá z nasledujúcich podmienok:

- uzol je čistý, t. j. obsahuje rovnaké hodnoty výstupnej premennej,
- všetky pozorovania v uzle majú rovnaké hodnoty vstupných premenných,
- strom dosiahol používateľom definovanú hĺbku vetvenia,
- počet pozorovaní v rodičovskom uzle je menší ako používateľom definovaný minimálny počet pozorovaní,
- počet pozorovaní v dcérskych uzloch je menší ako používateľom definovaná minimálna hranica,
- redukcia nečistoty uzla, ktorý by sa mal optimálne rozštiepiť, je nižšia, ako ju používateľ definoval.

[6] *Pri inkrementálnych úlohách sa postupne spracúva jeden trénovací prípad za druhým. Po každom prípade použitý algoritmus poskytuje riešenie. Pri neinkrementálnych úlohách sa spracuje naraz celá množina prípadov.*

Pri generovaní rozhodovacích stromov vedie snaha o podrobný opis údajov trénovacej množiny k vytvoreniu stromu, ktorý bezchybne klasifikuje na množine trénovacích prípadov. Takýto strom býva často preučený. Preučenie (*overfitting*) stromu v praxi znamená, že model síce kvalitne vysvetľuje vzťahy na trénovacej množine, tie však nie sú všeobecne platné, preto pri jeho aplikácii na inej množine údajov nastáva vysoká chybovosť. Typickými znakmi preučeného stromu sú jeho prílišná košatosť, tenké vetvy obsahujúce často iba jeden tréningový prípad a málopočetné listové uzly na spodných úrovniach stromu.

Vygenerované stromy sú preto modifikované tzv. orezávaním (*tree-pruning*). Používajú sa dva spôsoby orezávania:

- orezávanie pri konštrukcii (*prepruning*),
- orezávanie po konštrukcii (*postpruning*).

Pri prvom spôsobe orezávania počas rastu stromu sa predčasne pomocou modifikovaného algoritmu ukončí rast niektorých vetiev. Dôvodom ukončenia môže byť napríklad dostatočne vysoká pravdepodobnosť, že údaje príslušnej vetvy patria do tej istej klasifikačnej triedy. V praxi môže byť problémom určenie tejto hranice pravdepodobnosti.

Ďalšou možnosťou, ako zlepšiť predikčnú schopnosť a chybu klasifikácie stromu, je orezanie už vygenerovaného stromu. Pri tomto spôsobe sa vygeneruje úplný strom. Pri prerezávaní, keď sa nelistové uzly nahrádzajú listovými, sa posudzuje, ako sa zhorší jeho klasifikačná schopnosť. Tento spôsob sa v praxi považuje za jednoduchší, hodnovernejší, hoci je časovo náročnejší.

Techniky orezávania sa líšia od seba aj tým, aká množina údajov sa pri raste stromu a jeho spätnom orezaní používa. Techniky orezávania podľa toho rozdeľujeme do dvoch skupín:

- orezávanie používajúce len trénovaciu množinu,
- orezávanie používajúce trénovaciu aj testovaciu, resp. validačnú množinu.

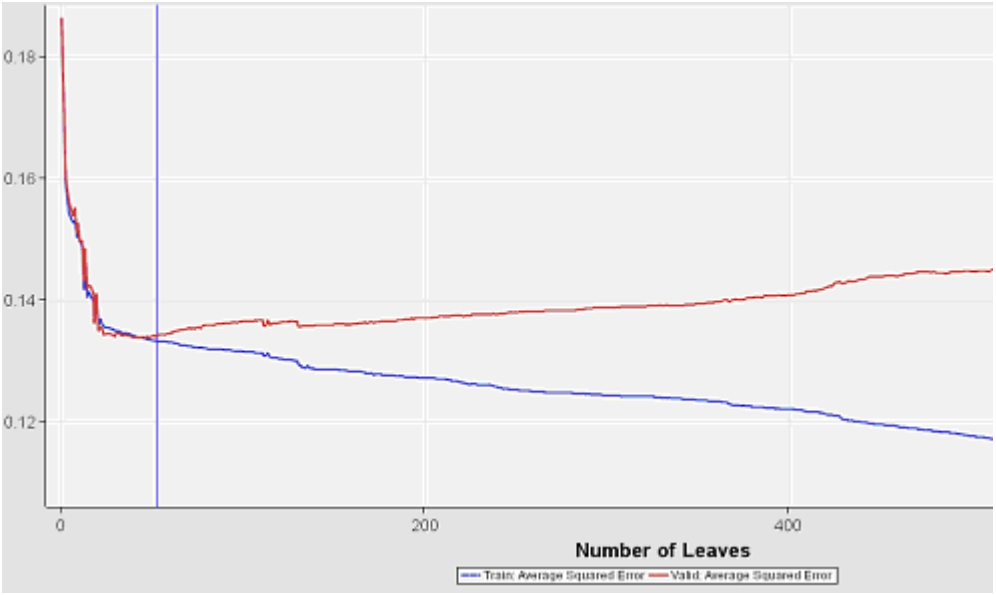
Pri prvej technike sa na tej istej dátovej množine, na ktorej sa nechá strom narásť, robí rozhodnutie o tom, ako ho orezať.

Pri druhej technike sa na jednej (trénovacej) množine strom vygeneruje, druhá množina slúži na výber podstromu z množiny všetkých kandidátskych podstromov, ktorým je pôvodný strom nahradený. Pri výbere sa zohľadňuje miera nesprávnej klasifikácie.[7]

Na obrázku 3 je znázornená závislosť priemernej štvorcovej chyby od počtu listov pri raste stromu. S rastúcim počtom listov priemerná štvorcová chyba na trénovacej množine systematicky klesá, na validačnej množine začne od istého počtu listov rásť. Problém sa rieši orezaním stromu na taký počet listov, aby chyba na trénovacej aj validačnej množine dosiahla minimum (na obrázku je orezanie naznačené zvislou čiarou).

[7] *Postup orezávania podrobnejšie pozri v [2] na s. 184 – 192.*

Obrázok č. 3: Orezanie rozhodovacieho stromu



Zdroj: vlastné spracovanie

5. PRAKTICKÁ UKÁŽKA ROZHODOVACIEHO STROMU

Ukážkou rozhodovacieho stromu, ktorý bol vytvorený na reálnej databáze údajov, je klasifikačný strom vytvorený na základe údajov pochádzajúcich zo štatistického zisťovania EU SILC 2015 (zdroj: ŠÚ SR, EU SILC 2015, UDB 26/09/2016). Použili sme R_súbor (Register osôb), ktorý obsahuje záznam za každú osobu, ktorá v čase zisťovania žila v domácnosti zahrnutej do databázy alebo bola dočasne neprítomná. Tento súbor obsahoval 16 181 prípadov (osôb). Databáza bola rozdelená na trénovaciu množinu (60 % prípadov) a validačnú množinu (40 % prípadov).

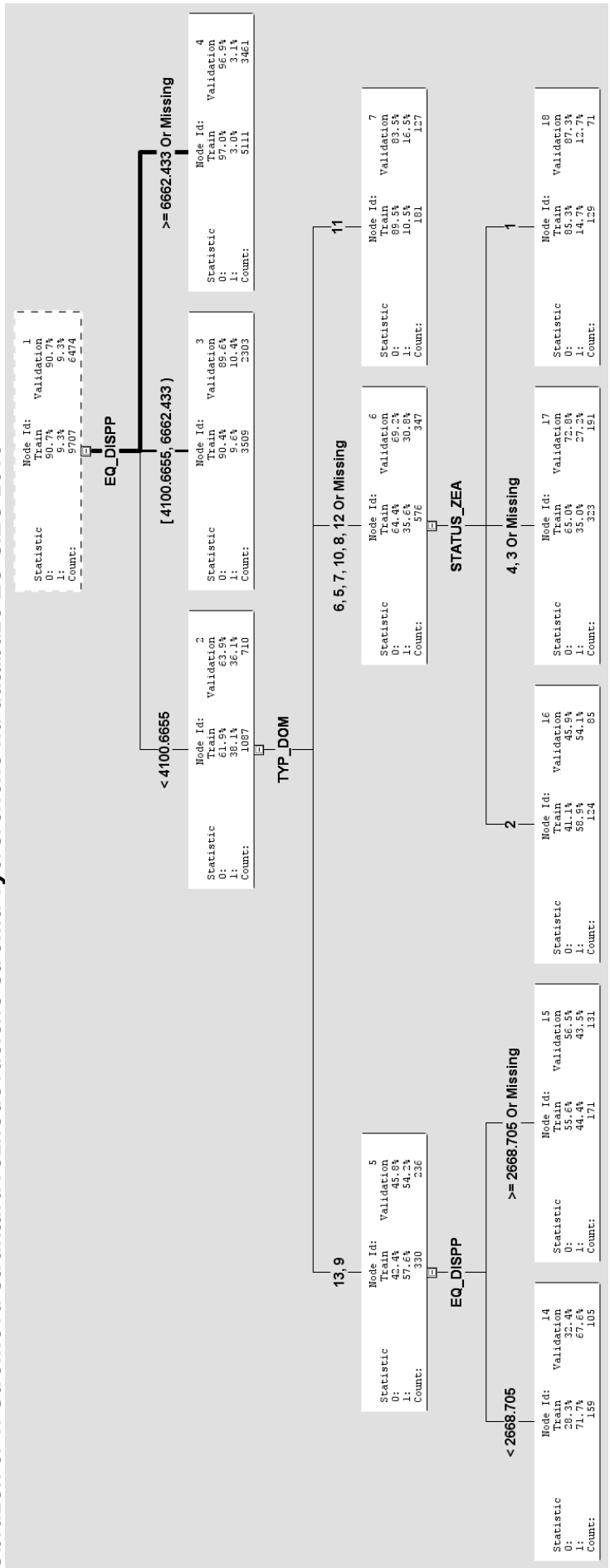
Na modelovanie sme definovali skupinu vstupných (vysvetľujúcich, nezávislých) premenných a modelovanú (vysvetľovanú, závislú) premennú. Nezávislými premennými boli premenné: **RB090**: *pohlavie,* (1 – muž, 2 – žena), **RB210**: *status základnej ekonomickej aktivity* (1 – pracujúci, 2 – nezamestnaný, 3 – starobný dôchodca, osoba v predčasnom dôchodku, 4 – iná neaktívna osoba), **HT** *typ domácnosti* (5 – jednočlenná domácnosť, 6 – domácnosť 2 dospelých bez závislých detí – obaja vo veku pod 65 rokov, 7 – domácnosť 2 dospelých bez závislých detí – aspoň jeden dospelý vo veku 65 rokov a viac, 8 – ostatné domácnosti bez závislých detí, 9 – domácnosť s 1 rodičom a s 1 alebo viac závislými deťmi, 10 – domácnosť 2 dospelých s 1 závislým dieťaťom, 11 – domácnosť 2 dospelých s 2 závislými deťmi, 12 – domácnosť 2 dospelých s 3 alebo viac závislými deťmi, 13 – ostatné domácnosti so závislými deťmi), **EQ_INC20**: ekvivalentný disponibilný príjem domácnosti (ročná suma). Závislou premennou bola premenná **SEV_DEP**: *závažná materiálna deprivácia*[8] (1 – áno, 0 – nie), ktorá vyjadrovala, či je osoba alebo domácnosť ohrozená rizikom chudoby.

[8] *Za závažne deprivované osoby sa považujú osoby, ktoré uvádzajú neprítomnosť alebo vynútený nedostatok aspoň v štyroch z týchto deviatich položiek: čeliť neočakávaným výdavkom, v priebehu jedného roka jeden týždeň dovolenky mimo domova, platiť za nedoplatky (hypotéky alebo nájomné, účty alebo kúpy na splátky), jedlo s mäsom, hydinou alebo rybou každý druhý deň, udržiavať primerane vykurovaný domov. Ďalšie položky vyjadrujú, že domácnosť si nemôže dovoliť (hoci chce): vlastniť práčku; vlastniť farebný televízor; vlastniť telefón; vlastniť osobný automobil. Išlo o predmety dlhodobej spotreby alebo činnosti, ktorých neprítomnosť, resp. nedostatok boli vynútené (ľudia by to chceli vlastniť, ale zdroje im to nedovoľujú).*

Klasifikačný strom sme vytvorili v programe SAS Enterprise Miner™, ktorý využíva vlastnú metodológiu SEMMA. Ako mieru čistoty údajov sme zvolili entropiu, umožnili sme najviac trojnásobné vetvenie, rast stromu sme regulovali tým, že sme povolili maximálne štyri úrovne vetvenia a určili sme minimálny počet prípadov v listoch. Vytvorený klasifikačný strom je na obrázku č. 4.

Najväčší vplyv na závažnú materiálnu depriváciu má výška ekvivalentného disponibilného príjmu, vplyv pohlavia je zanedbateľný. Najvyšší podiel materiálne deprivovaných je v skupine osôb, u ktorých výška ekvivalentného disponibilného príjmu neprekračuje 4 101 eur, pričom v tejto skupine sa podiel materiálne deprivovaných líši v závislosti od typu domácnosti, v ktorej osoba žije. Napríklad v kategórii domácností 2 dospelých s 2 závislými deťmi je 16,5 % deprivovaných (údaje validačnej množiny), v kategórii domácností s 1 rodičom a s 1 alebo viac závislými deťmi a ostatných domácností so závislými deťmi je až 54,2 % materiálne deprivovaných. V prípade, že výška ich ročného ekvivalentného disponibilného príjmu klesne pod 2 668,70 eura, zvýši sa podiel materiálne deprivovaných na 67,6 % (údaje na validačnej množine). Ak by sme rozdelili osoby s výškou ročného ekvivalentného disponibilného príjmu pod hranicou 4 101 eur žijúce v ostatných domácnostiach (*typ domácnosti*: 5, 6, 7, 8, 10 a 12) podľa statusu základnej ekonomickej aktivity, najvyšší podiel materiálne deprivovaných by bol v skupine nezamestnaných (54,1 % na validačnej množine).

Obrázok č. 4: Stromová štruktúra rozhodovacieho stromu vytvoreného na databáze EU SILC 2015



Poznámka: EQ_DISPP – ekvivalentný disponibilný príjem domácnosti, TYP_DOM – typ domácnosti, STATUS_ZEA – status základnej ekonomickej aktivity.
Zdroj údajov: ŠÚ SR, EU SILC 2015, UDB 26/09/2016, vlastné spracovanie (SAS Enterprise Miner™)

6. ZÁVER

V štatistickej praxi sa často stretávame so situáciami, keď je našou úlohou analyzovať dátové súbory vyznačujúce sa vysokou dimenzionalitou, súvisiacou so snahou komplexne opísať zložený jav, pričom premenné, ktoré tento jav opisujú, môžu mať rôzny charakter (nominálne, ordinálne, kardinálne). Rozhodovacie stromy (klasifikačné, regresné stromy) patria k viacrozmerným metódam, ktoré sú schopné analyzovať takéto dátové súbory. Rozhodovacie stromy možno považovať za alternatívny nástroj k regresnej analýze (v situáciách, keď je závislá premenná číselná spojitá); používajú sa ako alternatíva k logistickej regresii a diskriminačnej analýze (ak je modelovaná premenná kategoriálna). Niekedy sa táto metóda zaraďuje do skupiny hierarchických zhlukovacích metód, pretože jej produktom je vytvorenie disjunktných podskupín z pôvodného súboru objektov [5].

V článku sme sa venovali opisu rozhodovacích stromov, metódam výberu premenných (testovacích znakov) na vetvenie a stručnému opisu najčastejšie používaných algoritmov tvorby rozhodovacích stromov.

Článok vznikol s podporou grantovej agentúry VEGA v rámci projektu VEGA 1/0770/17 Dostupnosť bývania na Slovensku.

LITERATÚRA

[1] BERKA, P.: Dobývání znalostí z databází. Praha: Academia, 2003. ISBN 80-200-1062-9.
[2] BERRY, M. J. A. – LINOFF, G. S.: Data mining Techniques. For Marketing, Sales, and Customer Relationship management. Indianapolis: Wiley Publishing, Inc., 2004.
[3] BIGGS, D. – DE VILLE, B. – SUEN, E.: A method of choosing multiway partitions for classification and decision trees. In: Journal of Applied Statistics [online], 1991, No. 1, p. 49-62. Dostupné na internete: <http://dx.doi.org/10.1080/02664769100000005> [prístup k 11. 2. 2017].
[4] BREIMAN, L. – FRIEDMAN, J. H. – OLSHEN, R. A. – STONE, C. J.: Classication and Regression Trees. Wadsworth, 1984. ISBN 9780412048418.
[5] HENDL, J.: Přehled statistických metod: Analýza a metaanalýza dat. Praha: Portál, 2009. ISBN 978-80-7367-482-3.
[6] KASS, G. V.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. In: Applied Statistics, 1980, No. 2, p. 119-127.
[7] KLASCHKA, J. – KOTRČ, E.: Klasifikační a regresní lesy. In: Sborník konference ROBUST [online]. Jednota českých matematiků a fyziků, 2004, s. 177 – 184 Dostupné na internete: <http://statspol.cz/robust/robust2004/klaschka.pdf> [prístup k 15. 2. 2017].
[8] MACHOVÁ, K.: Strojové učenie: Princípy a algoritmy. Košice, 2002.
[9] MORGAN, J. N. – MESSENGER, R. C.: THAID: a sequential program for the analysis of nominal scale depend variables [online]. University of Michigan, 1973. Dostupné na internete: <http://hdl.handle.net/2027/mdp.39015071883859> [prístup k 17. 2. 2017].
[10] PARALIČ, J.: Umelá inteligencia 1: Objavovanie znalostí [online]. Dostupné na internete: <http://people.tuke.sk/jan.paralic/prezentacie/UI/objavovanie_znalosti.pdf> [prístup k 17. 2. 2017].
[11] QUINLAN, J. R.: Induction of Decision Trees. In: Machine Learning [online], 1986, No. 1, p. 81-106. Dostupné na internete: <https://doi.org/10.1007/BF00116251> [prístup k 17. 2. 2017].

[12] QUINLAN, J. R.: Simplifying decision trees. In: International Journal of Man-Machine Studies [online], 1987, No. 3, p. 221-234. Dostupné na internete: <https://doi.org/10.1016/S0020-7373(87)80053-6> [prístup k 17. 2. 2017].
[13] QUINLAN, J. R.: C4.5 Programs for Machine Learning. San Francisco: Morgan Kaufmann, 1993. ISBN 1-55860-238-0.
[14] ŘEZANKOVÁ, H.: Analýza dat z dotazníkových šetření. Praha: Professional Publishing, 2007. ISBN 978-80-86946-49-8.
[15] SONQUIST, J. N. – MORGAN, J. A.: Problems in the Analysis of Survey Data, and a Propsal. In: Journal of the American Statistical Association, 1963, No. 302, p. 415-435.
[16] STANKOVIČOVÁ, I.: Rozhodovacie stromy v marketingových analýzach. In: Nová ekonomika, 2006, č. 1, s. 105 – 111.
[17] TEREK, M. – HORNÍKOVÁ, A. – LABUDOVÁ, V.: Hĺbková analýza údajov. Bratislava: Iura Edition, 2010. ISBN 978-80-8078-336-5.
[18] WILKINSON, L.: Tree Structured Data Analysis: AID, CHAID and CART [online]. Dostupné na internete: http://www.spss.com/research/wilkinson/Publications/c&r trees.pdf [prístup k 17. 2. 2017].

RESUME

The tree based learning algorithms are considered to be one of the best and most common supervised learning methods. They are suitable for exploratory data analysis in obtaining information on the impact of the large number of candidate input variables on the target variable. Decision tree models can be effectively used to determine the most important attributes in a dataset. Decision tree is a data mining technique used for the classification and the prediction of values. In data mining, it represents classifications and regression models. When a decision tree is used for classification tasks (the target variable is categorical), it is referred to as a classification tree. When it is used for regression tasks (the target variable is continuous), it is called a regression tree. Decision trees have many appealing properties: understandability, flexibility in handling a variety of input data: nominal, numeric and textual, adaptability in processing datasets with error or missing values, achieving high predictive performance for a relatively small computational effort, being part of various data mining software. Because decision trees combine both data exploration and modeling techniques, they are a powerful first step in the modeling process even in the construction of the final model using some other technique. This article describes the structure of decision trees and the basic algorithm for their construction.

PROFESIJNÝ ŽIVOTOPIS

Doc. RNDr. Viera Labudová, PhD., *je absolventkou Matematicko-fyzikálnej fakulty Univerzity Komenského v Bratislave. Na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave pôsobila od roku 2000 ako odborná asistentka, od roku 2014 vo funkcii docentky v študijnom odbore kvantitatívne metódy v ekonómii. Vo svojej vedeckovýskumnej a pedagogickej činnosti sa venuje aplikácii štatistických metód pri analýzach sociálno-ekonomických javov, analýzam sociálno-patologických javov s osobitným zreteľom na výskyt chudoby, aplikácii metód hĺbkovej analýzy údajov, analýze kategoriálnych údajov a regionálnej štatistike.*

KONTAKT

viera.labudova@euba.sk