SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2020
ročník/volume 30

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

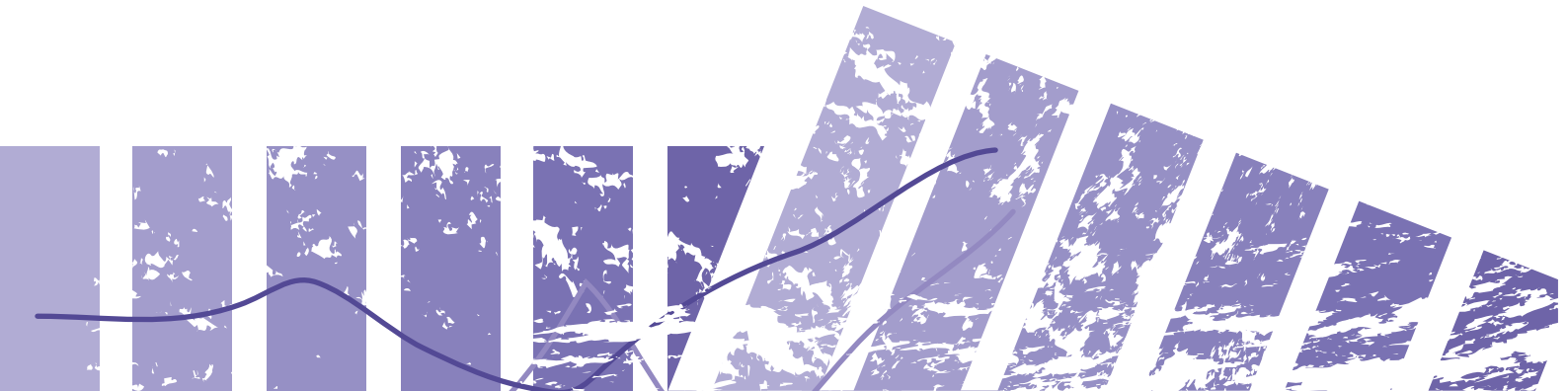
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

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

Strany/Pages: 40 – 53

Dátum vydania/Publication date: 15. júl 2020/July 15, 2020



Hana ŘEZANKOVÁ
Fakulta informatiky a statistiky Vysoké školy ekonomické v Praze

ZPŮSOBY VÝBĚRU VYSVĚTLUJÍCÍCH PROMĚNNÝCH V KLASIFIKAČNÍCH STROMECH

METHODS OF SELECTING EXPLANATORY VARIABLES IN CLASSIFICATION TREES

ABSTRAKT
Článek se zaměřuje na různá hodnocení vztahů mezi kategoriálními proměnnými a jejich aplikaci na problematiku výběru vysvětlujících proměnných v klasifikačních stromech. Jsou jednak diskutovány postupy dostupné v komerčních programových systémech (chí-kvadrát testy a porovnávání variability vysvětlované proměnné v různých skupinách objektů pomocí Giniho míry a entropie), jednak naznačeny další možnosti vývoje v této oblasti. Stávající postupy jsou ilustrovány na analýze dat v programovém systému IBM SPSS Decision Trees. Výzkum se v poslední době zaměřuje na hodnocení jednostranné závislosti ordinální vysvětlované proměnné na proměnné nominální a implementaci v rámci klasifikačních stromů. Takové přístupy již byly realizovány v balíčcích v prostředí R.

ABSTRACT
The paper focuses on different evaluations of relationships between categorical variables and their application to the explanatory variables selection in classification trees. On the one hand approaches available in commercial software systems (chi-square tests and comparison of variability explained in different groups of objects using the Gini measure and the entropy) are discussed and on the other hand, further development possibilities of development are outlined. The well-known possibilities are illustrated on the data analysis in the IBM SPSS Decision Trees system. Recently research focuses on the evaluation of directional association of the target variable on the nominal variable and the implementation in classification trees. Such approaches have been realized in the packages in the R environment.

KLÍČOVÁ SLOVA
klasifikační stromy, kategoriální proměnná, nominální proměnná, ordinální proměnná

KEY WORDS
classification trees, categorical variable, nominal variable, ordinal variable

1. ÚVOD
Názvem *klasifikační stromy* je označována skupina metod, které byly navrženy pro řešení klasifikačních úloh s vysvětlovanou proměnnou. V klasifikačních úlohách tohoto typu jsou na základě známých hodnot kategoriální vysvětlované proměnné s využitím vysvětlujících proměnných vytvářeny modely či pravidla tak, aby mohly být odhadovány hodnoty vysvětlované proměnné v případě, kdy nejsou známy. Cílem je získání návodu k zařazování (klasifikaci) objektů charakterizovaných vektory hodnot vysvětlujících proměnných do skupin (tříd) daných množinou kategorií vysvětlované proměnné.

Klasifikační stromy tedy slouží ke stejným účelům jako diskriminační analýza nebo logistická regrese. Jejich základní odlišností od dvou dalších zmíněných metod je to, že vysvětlující proměnné jsou uvažovány jako kategoriální. Buď do analýzy vstupují jako kategoriální, nebo jsou na kategoriální převedeny (v případě vstupních kvantitativních spojitých proměnných). Navíc v procesu analýzy může docházet k překódování vysvětlujících proměnných, aby výsledné vztahy byly co nejvýstižnější.

Základním principem klasifikačních stromů je postupný výběr vysvětlujících proměnných z množiny vstupních proměnných (v případě potřeby překódovaných do vhodného počtu kategorií). Je vytvářena stromová hierarchická struktura, při které je původní soubor objektů postupně rozdělován na podsoubory. Není ovšem vhodné označovat klasifikační stromy jako metodu hierarchické shlukové analýzy (jak se někdy mylně v literatuře uvádí), neboť shluková analýza klasifikuje objekty na zcela jiném principu, a to bez využití reálné vysvětlované proměnné, navíc ani není znám počet tříd, viz [9].

Přestože jsou metody pro tvorbu klasifikačních stromů v dnešní době již poměrně dobře známé, v literatuře se někdy vyskytují nepřesnosti. Cílem tohoto článku je diskutovat některé způsoby výběru vysvětlujících proměnných. Nebude zde detailně pojednáno o konstrukci stromové struktury, ani o možných podmínkách ukončení větvení stromu. Základy těchto postupů jsou uvedeny např. v článku [7], kde jsou též charakterizovány nejznámější metody, včetně jejich historie. Budou však naznačeny další možnosti vývoje v této oblasti, k nimž patří např. aplikace speciálních postupů pro ordinální vysvětlovanou proměnnou.

2. KRITÉRIA PRO VÝBĚR VYSVĚTLUJÍCÍCH PROMĚNNÝCH

Při výběru vysvětlujících proměnných je postupně pro všechny vstupní proměnné (s různými vhodnými počty kategorií) buď testována nezávislost mezi vysvětlovanou a vysvětlující proměnnou, nebo je posuzována vnitroskupinová, resp. meziskupinová variabilita vysvětlované proměnné při rozdělení objektů do skupin podle kategorií zkoumané vysvětlující proměnné. Pokud jsou vysvětlující proměnné ordinální, pak se pořadí kategorií zohledňuje pouze při překódování do menšího počtu kategorií. Při samotném výběru vysvětlujících proměnných jsou pak všechny vstupní proměnné považovány za nominální. Většina používaných postupů je vhodná pro nominální vysvětlovanou proměnnou.

2.1. VYUŽITÍ CHÍ-KVADRÁT TESTŮ

Chí-kvadrát testy jsou určeny pro zkoumání závislosti dvou nominálních proměnných; v klasifikačních stromech se zpravidla používají bez ohledu na typ kategoriálních proměnných. Při jejich aplikaci se na všech vytvářených úrovních, tzn. pro různé skupiny objektů, provádějí testy o nezávislosti vysvětlované proměnné postupně s jednotlivými vysvětlujícími proměnných, přičemž pro vícekategoriální proměnné (tj. proměnné s více než dvěma kategoriemi) jsou prováděna překódování do nových proměnných s různými počty kategorií (původní kategorie jsou různými způsoby sdružovány). Pro větvení stromu se vybírá vysvětlující proměnná (původní nebo překódovaná), pro kterou je p-hodnota menší nebo rovna stanovené hladině významnosti. Pokud je takových proměnných více, vybírá se proměnná s nejnižší p-hodnotou. Protože jde o opakované statistické testování,

p-hodnota je obvykle modifikována Bonferroniho metodou. Pokud je p-hodnota větší než stanovená hladina významnosti, strom se dále nevětví.
Chí-kvadrát test může být proveden buď pomocí Pearsonovy chí-kvadrát statistiky, nebo s využitím věrohodnostního poměru. Označme vysvětlovanou proměnnou jako Y a její kategorie y_j, kde $j = 1, 2, \ldots, s$, a zkoumanou vysvětlující proměnnou jako X s kategoriemi x_i, kde $i = 1, 2, \ldots, r$. Dále označme počet objektů v rozdělované skupině symbolem n, sdružené absolutní četnosti v dvourozměrné kontingenční tabulce pro danou skupinu jako n_{ij}, řádkové marginální četnosti jako n_{i+} a sloupcové marginální četnosti jako n_{+j}. Pearsonova chí-kvadrát statistika se počítá jako:

$$\chi_P^2 = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}. \qquad (1)$$

Při platnosti nulové hypotézy o nezávislosti má tato statistika přibližně chí-kvadrát rozdělení s počty stupňů volnosti $(r-1)(s-1)$. Vzorec pro věrohodnostní poměr je:

$$\chi_{LR}^2 = -2\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}\ln\left(\frac{\frac{n_{i+}n_{+j}}{n}}{n_{ij}}\right). \qquad (2)$$

Při platnosti nulové hypotézy o nezávislosti má tato statistika rovněž přibližně stejné rozdělení jako Pearsonova statistika.

2.2. Využití principu analýzy rozptylu pro nominální vysvětlovanou proměnnou

Jinou z používaných technik je nalezení vždy takové vysvětlující proměnné, pomocí jejichž kategorií lze vytvořit podmnožiny objektů tak, aby vnitroskupinová variabilita vysvětlované proměnné byla co nejmenší. Jde tedy o aplikaci poznatků z analýzy rozptylu. Protože je ale vysvětlovaná proměnná kategoriální, místo součtů čtvercových odchylek (jako základu pro výpočet rozptylu) se používají speciální míry pro nominální proměnnou, a to buď nominální rozptyl (Giniho míra mutability, viz [5]), nebo entropie. Ve statistických programových systémech IBM SPSS Decision Trees a Statistica je nabízena pouze Giniho míra. Využijme symboliku z části 2.1 s tím rozdílem, že n bude vždy vyjadřovat celkový počet objektů. Za předpokladu, že vysvětlovaná proměnná je v kontingenční tabulce sloupcová, pak celkovou variabilitu proměnné Y lze pomocí Giniho míry vyjádřit jako:

$$G(Y) = \sum_{j=1}^{s} \frac{n_{+j}}{n}\left(1 - \frac{n_{+j}}{n}\right), \qquad (3)$$

přičemž podíly $\frac{n_{+j}}{n}$ jsou marginální relativní četnosti, které lze označit též jako p_{+j}. Tato míra zohledňuje podíl počtu párů objektů s různými hodnotami. Nabývá hodnot z intervalu

od 0 do $(s-1)/s$; 0 odpovídá konstantě, nejvyšší hodnota pak stejným četnostem pro všechny kategorie, viz [12] a [13].

Při rozdělení původní množiny objektů na základě kategorií vybrané vysvětlující proměnné X pak lze vyjádřit variabilitu vysvětlované proměnné pro každou podmnožinu objektů (tj. pro každý řádek v kontingenční tabulce). Vnitroskupinová variabilita je váženým průměrem z hodnot Giniho míry získaných pro všechny podmnožiny objektů, tj.:

$$G(Y|X) = \sum_{i=1}^{r} \frac{n_{i+}}{n} \sum_{j=1}^{s} \frac{n_{ij}}{n_{i+}} \left(1 - \frac{n_{ij}}{n_{i+}}\right). \tag{4}$$

Při prvním větvení stromu se porovnává celková variabilita s vnitroskupinovou variabilitou rozdílem, to znamená, že je spočtena meziskupinová variabilita jako

$$G(Y) - G(Y|X). \tag{5}$$

Pro větvení se vybere taková vysvětlující proměnná, pro kterou byla zjištěna největší meziskupinová variabilita (tudíž nejmenší vnitroskupinová variabilita).

Při dalším větvení je postupováno analogicky. Porovnává se vážená variabilita proměnné Y v určité skupině objektů a vnitroskupinová variabilita při rozdělení dané skupiny objektů do dílčích skupin. Váhy se počítají vždy jako podíl počtu objektů v dané skupině k celkovému počtu objektů n. Obecně lze vzorec pro variabilitu proměnné Y v množině objektů odpovídající u-tému uzlu zapsat jako:

$$G_u(Y) = \frac{n_u}{n} \sum_{j=1}^{s} \frac{n_{+ju}}{n_u} \left(1 - \frac{n_{+ju}}{n_u}\right), \tag{6}$$

kde symbol u označuje uzel, v kterém je prováděno větvení. Např. n_u označuje počet objektů v u-tém uzlu a platí, že $n_0 = n$.

Vnitroskupinovou variabilitu pro u-tý uzel vyjádříme opět jako vážený průměr z hodnot Giniho míry získaných pro všechny podmnožiny objektů, tj.:

$$G_u(Y|X) = \sum_{i=1}^{r} \frac{n_{i+u}}{n} \sum_{j=1}^{s} \frac{n_{iju}}{n_{i+u}} \left(1 - \frac{n_{iju}}{n_{i+u}}\right). \tag{7}$$

Meziskupinová variabilita pro u-tý uzel je dána vztahem:

$$G_u(Y) - G_u(Y|X). \tag{8}$$

Obdobně by mohla být vyjádřena variabilita vysvětlované proměnné Y pomocí entropie. V metodách statistické analýzy se entropie vyjadřuje pomocí přirozeného logaritmu, který je využit i v jiných postupech (viz např. výše uvedený věrohodnostní poměr). V metodách

„data mining“ se využívá dvojkový logaritmus. Tento postup je v klasifikačních stromech využit např. v systému SAS Enterprise Miner, viz příklad v [7].

Celkovou variabilitu proměnné Y lze pomocí entropie vyjádřit jako:

$$H(Y) = \sum_{j=1}^{s} \frac{n_{+j}}{n} \log_2 \left(\frac{n_{+j}}{n}\right). \tag{9}$$

Při rozdělení původní množiny objektů na základě kategorií vybrané vysvětlující proměnné X pak lze vyjádřit variabilitu vysvětlované proměnné Y pro každou podmnožinu objektů. Vnitroskupinová variabilita je váženým průměrem z hodnot získaných pro všechny podmnožiny objektů, tj.:

$$H(Y|X) = \sum_{i=1}^{r} \frac{n_{i+}}{n} \sum_{j=1}^{s} \frac{n_{ij}}{n_{i+}} \log_2 \left(\frac{n_{ij}}{n_{i+}}\right). \tag{10}$$

Při prvním větvení stromu se porovnání celková variabilita s vnitroskupinovou variabilitou rozdílem, to znamená, že je spočtena meziskupinová variabilita jako:

$$H(Y) - H(Y|X). \tag{11}$$

Výsledná hodnota je označována jako „informační zisk“ (viz [7]). Při dalším větvení je postupováno analogicky s postupem vysvětleným při aplikaci Giniho indexu. Celková variabilita určité skupiny je vždy vážena relativní četností objektů vzhledem k celkovému původnímu počtu objektů.

Variabilita nominální vysvětlované proměnné a její rozklad jsou v praxi aplikovány také ke konstrukci koeficientů jednostranné závislosti (na jiné nominální proměnné), pro které byly rovněž navrženy testy na nulovost těchto koeficientů, tj. nezávislost. Jednostranná závislost je posuzována na základě podílu meziskupinové variability na celkové variabilitě. Podle způsobu vyjádření variability nominální proměnné jsou v praxi využívány koeficienty Goodmanovo-Kruskalovo lambda, Goodmanovo-Kruskalovo tau (využívá Giniho míru) a koeficient neurčitosti či nejistoty neboli informační koeficient (využívá entropii vyjádřenou pomocí přirozeného logaritmu), viz např. [13]. Tyto postupy však nejsou v klasifikačních stromech implementovány.

Využití měr variability nominální proměnné v analýze dat je samozřejmě mnohem širší, jako příklad lze uvést konstrukce nových měr podobnosti, které mohou být aplikovány ve shlukové analýze kategoriálních dat, viz [14].

2.3. POSTUPY PRO ORDINÁLNÍ VYSVĚTLOVANOU PROMĚNNOU

Pro ordinální vysvětlovanou proměnnou je obvykle postupováno analogicky jako pro nominální proměnnou, ovšem s využitím vhodného testu (v systému IBM SPSS Decision Trees je nabízen pouze chí-kvadrát test s využitím věrohodnostního poměru) či vhodné míry variability pro ordinální proměnnou (v systému SAS Enterprise Miner jsou speciálně

upraveny výpočty pro Giniho míru a entropii). Existují ale i další přístupy vyjádření míry závislosti ordinální proměnné na nominální, viz např. [1].

Variabilita ordinální proměnné je obvykle vyjadřována pomocí míry známé pod označením *dorvar* (diskrétní ordinální variance), viz např. [12]. Za předpokladu využití v kontingenční tabulce ji můžeme zapsat jako:

$$dorvar(Y) = 2\sum_{j=1}^{s} F_{+j}\left(1 - F_{+j}\right), \tag{12}$$

kde F_{+j} je marginální kumulativní relativní četnost pro j-tou kategorii proměnné Y. Tato míra nabývá hodnot od 0 do $(s-1)/2$. Variabilita se zvyšuje se vzrůstajícími četnostmi v krajních kategoriích (v první a poslední kategorii), viz [13]. Vztah této míry k Giniho míře mutability je vysvětlen v [10], kde je rovněž navržena míra závislosti charakterizující závislost ordinální (vysvětlované) proměnné na proměnné nominální (vysvětlující). Podrobněji se měřením variability ordinální proměnné zabývá článek [3], v němž je kromě řady jiných měr zmíněna normalizovaná varianta míry *dorvar* (vyjádřená na intervalu od 0 do 1). Vyjádřením variability ordinální proměnné a jeho využitím při ohodnocení závislosti se zabývá též článek [8].

Možné využití vyjádření variability ordinální proměnné pomocí kumulativních relativních četností k výběru vysvětlujících proměnných při konstrukci klasifikačních stromů je navrženo v článku [11]. Na návrhy publikované v tomto článku navazuje Archer, který v prostředí R vytvořil balíček rpartOrdinal, viz [2]. Další autoři navrhli modifikaci Archerova balíčku a vytvořili balíček rpartScore, viz [4]. Problematika ordinální vysvětlované proměnné při konstrukci náhodných lesů (rozšíření problematiky klasifikačních stromů pro případy datových souborů o velké dimenzionalitě) je zohledněna např. v [6].

Giniho míra mutability a míra *dorvar* mohou být vyjádřeny zobecněným vzorcem, jehož speciálním případem je rovněž míra pro diskrétní kvantitativní vysvětlovanou proměnnou. Je to Giniho průměrná diference (mean difference) daná vztahem:

$$D(Y) = 2\sum_{j=1}^{s-1}\left(y_{j+1} - y_j\right)F_{+j}\left(1 - F_{+j}\right), \tag{13}$$

odvození viz [10]. Pro kvantitativní proměnnou je však možné použít také rozptyl.

3. ILUSTRACE APLIKOVÁNÍ VYBRANÝCH KRITÉRIÍ

Pro účely ilustrace výše uvedených postupů je vybrán jednoduchý příklad s 11 objekty a třemi vysvětlujícími proměnnými. Objekty (tj. statistické jednotky pro klasifikační stromy) jsou vybrané metody shlukové analýzy ze tří skupin, kterými jsou *metody pevného rozkladu*, *metody fuzzy rozkladu* a *metody hierarchického shlukování* (vysvětlovaná proměnná *skupina*), viz tabulka č. 1. První vysvětlující proměnnou je *shlukování*, která indikuje, zda jde o shlukování *pevné* (každý objekt je přiřazen právě do jednoho shluku),

nebo *fuzzy* (každé kombinaci objekt a shluk je přiřazen stupeň příslušnosti na škále od 0 do 1). Podle této proměnné je možné jednoznačně identifikovat metody fuzzy rozkladu. Druhá proměnná *centroid* nabývá tří kategorií podle toho, zda jsou v průběhu analýzy pro jednotlivé shluky vytvářeny centroidy (vektory charakteristik vstupních proměnných) jako *vektory průměrných hodnot* pro daný shluk, nebo jako *vektory mediánů*, nebo jestli *se centroidy nevytvářejí*. Třetí vysvětlující proměnná *vzdálenosti* charakterizuje, jaké typy vzdáleností jsou v průběhu analýzy počítány. Možnosti jsou *vzdálenosti objektů od centroidu*, *vzdálenosti objektů od medoidu* (konkrétní objekt ze souboru, který reprezentuje danou skupinu), *vzdálenosti mezi objekty z různých shluků* a *vzdálenosti mezi centroidy* (pro jednotlivé páry shluků).

K aplikaci klasifikačních stromů byl využit programový systém IBM SPSS Decision Trees (verze 26). Chí-kvadrát testy byly aplikovány v algoritmu CHAID, princip analýzy rozptylu pomocí Giniho míry v algoritmu CRT. U obou algoritmů byl nastaven malý minimální počet objektů v koncových uzlech (vzhledem k celkově malému počtu objektů v souboru); lze nastavit např. hodnoty 2 nebo 3. Jde o ilustraci používaných postupů na malém datovém souboru; v případě aplikace chí-kvadrát testů nejsou splněny podmínky pro jejich použití. Správně by měl být použit exaktní Fisherův test, příklad je proto pro srovnání doplněn p-hodnotami pro tento test.

Tabulka č. 1: Vstupní datová matice pro ilustraci výběru vysvětlujících proměnných

Metoda	*Shlukování*	*Centroid*	*Vzdálenosti*	*Skupina*
k-průměrů (HCM)	pevné	vektor průměrů	vzdálenosti objektů od centroidu	pevného rozkladu
k-mediánů	pevné	vektor mediánů	vzdálenosti objektů od centroidu	pevného rozkladu
k-medoidů (PAM)	pevné	nestanovuje se	vzdálenosti objektů od medoidu	pevného rozkladu
CLARA	pevné	nestanovuje se	vzdálenosti objektů od medoidu	pevného rozkladu
fuzzy *k*-průměrů (FCM)	fuzzy	vektor průměrů	vzdálenosti objektů od centroidu	fuzzy rozkladu
PCM	fuzzy	vektor průměrů	vzdálenosti objektů od centroidu	fuzzy rozkladu
fuzzy *k*-medoidů	fuzzy	nestanovuje se	vzdálenosti objektů od medoidu	fuzzy rozkladu
průměrného spojení	pevné	nestanovuje se	vzdálenosti mezi objekty z různých shluků	hierarchické metody
jednoduchého spojení	pevné	nestanovuje se	vzdálenosti mezi objekty z různých shluků	hierarchické metody
úplného spojení	pevné	nestanovuje se	vzdálenosti mezi objekty z různých shluků	hierarchické metody
centroidní	pevné	vektor průměrů	vzdálenosti mezi centroidy	hierarchické metody

Zdroj: vlastní zpracování

Z tabulky č. 1 je zřejmé, že lze jednoznačně identifikovat buď skupinu metod fuzzy rozkladu (na základě kategorie *fuzzy* proměnné *shlukování*), nebo skupinu hierarchických metod, která neobsahuje vzdálenosti objektů od centroidu, ani od medoidu, charakteristických pro metody rozkladu (proměnná *vzdálenosti*). Pomocí použitých klasifikačních stromů byly v různém pořadí vybírány právě proměnné s*hlukování* a *vzdálenosti*, jejichž kombinace vede k jednoznačnému přiřazení metod shlukové analýzy do stanovených skupin. V další části této kapitoly bude pro zjednodušení pozornost věnována těmto dvěma vysvětlujícím proměnným.

3.1. Aplikace chí-kvadrát testů

Při aplikaci chí-kvadrát testů se v prvním kroku provádějí testy o nezávislosti vysvětlované proměnné s jednotlivými vysvětlujícími proměnných, přičemž pro vícekategoriální proměnné jsou prováděna překódování do nových proměnných s různými počty kategorií. V tabulce č. 2 jsou uvedeny hodnoty získané pro Pearsonův chí-kvadrát test a Fisherův exaktní test pro vysvětlující proměnnou *Shlukování*, původní proměnnou *Vzdálenosti* a proměnné odvozené z této proměnné překódováním do dvou kategorií (výsledky pro proměnné vzniklé překódováním do tří kategorií nejsou pro zjednodušení uvedeny).

Tabulka č. 2: Hodnoty získané pro Pearsonův chí-kvadrát test a Fisherův exaktní test

	Pearsonova statistika	P-hodnota pro chí-kvadrát test	P-hodnota pro Fisherův test
Závislost na shlukování	11,000	0,004	0,006
Závislost na vzdálenostech (4 kategorie)	11,306	0,079	0,050
Závislost na vzdálenostech (2 kategorie – 1. varianta)	0,557	0,757	1,000
Závislost na vzdálenostech (2 kategorie – 2. varianta)	1,253	0,535	0,766
Závislost na vzdálenostech (2 kategorie – 3. varianta)	1,925	0,382	1,000
Závislost na vzdálenostech (2 kategorie – 4. varianta)	2,597	0,273	0,418
Závislost na vzdálenostech (2 kategorie – 5. varianta)	3,798	0,150	0,309
Závislost na vzdálenostech (2 kategorie – 6. varianta)	7,219	0,027	0,055
Závislost na vzdálenostech (2 kategorie – 7. varianta)	11,000	0,004	0,006

Zdroj: vlastní zpracování

Nejmenší p-hodnota je jak v případě Pearsonova testu (0,004), tak Fisherova testu (0,006) menší než 0,05 a je shodná pro dvě proměnné, kterými jsou *shlukování* a *vzdálenosti* překódované do dvou kategorií (7. varianta). Kontingenční tabulky pro dvě shodně nejlépe ohodnocené závislosti jsou uvedeny jako tabulky č. 3 a 4.

Tabulka č. 3: Kontingenční tabulka pro vztah skupiny metod a typu shlukování

		Skupina metod			Celkem
		metody pevného rozkladu	metody fuzzy rozkladu	hierarchické metody	
Typ shlukování	pevné	4	0	4	8
	fuzzy	0	3	0	3
Celkem		4	3	4	11

Zdroj: vlastní zpracování

Tabulka č. 4: Kontingenční tabulka pro vztah skupiny metod a výpočtu vzdálenosti

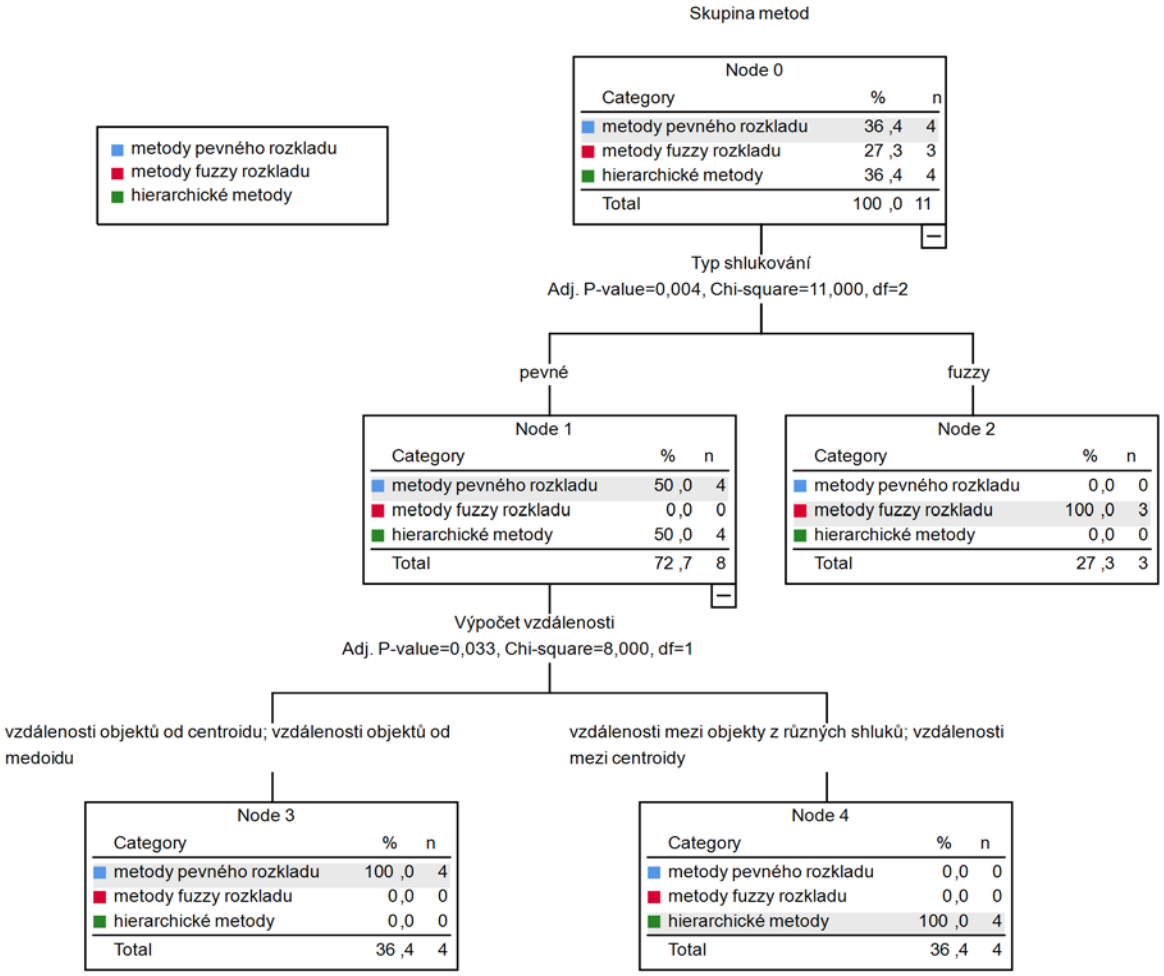
		Skupina metod			Celkem
		metody pevného rozkladu	metody fuzzy rozkladu	hierarchické metody	
Výpočet vzdálenosti	vzdálenosti objektů od centroidu nebo medoidu	4	3	0	7
	vzdálenosti mezi shluky	0	0	4	4
Celkem		4	3	4	11

Zdroj: vlastní zpracování

Algoritmem CHAID byla pro větvení použita první vysvětlující proměnná podle pořadí, tj. proměnná *shlukování*. Podle jejích dvou kategorií byly vytvořeny dvě podmnožiny objektů, viz obrázek č. 1. V každém uzlu grafu (anglicky „node“) je zobrazena tabulka četností pro hodnoty vysvětlované proměnné, které odpovídají dané skupině objektů. Do první skupiny byly zařazeny metody pevného shlukování, do druhé metody fuzzy shlukování. Druhá podmnožina tedy obsahuje pouze metody odpovídající kategorii „metody fuzzy rozkladu“ vysvětlované proměnné a další větvení se neprovádí. První podmnožina obsahuje metody ze dvou skupin, tudíž se zkoumá, zda by ji bylo možné dále rozdělit.

V úvahu přicházejí dvě zbývající vysvětlující proměnné a proměnné vytvořené jejich překódováním. Nejnižší p-hodnota byla získána při testu o nezávislosti vysvětlované proměnné s proměnnou *vzdálenosti* překódované do dvou kategorií, přičemž test byl proveden pro objekty z podmnožiny metod pevného shlukování. Odpovídající kontingenční tabulka pro dané dvě proměnné je označena jako tabulka č. 5. Hodnota Pearsonovy statistiky je 8, p-hodnota pak 0,005 a p-hodnota upravená Bonferroniho metodou 0,033. Jde o hodnotu menší než 0,05, proto se provádí další větvení stromu (při použití Fisherova testu by byla získána p-hodnota 0,014). Tím byly získány další dvě podmnožiny objektů, které jednoznačně odpovídají zbývajícím skupinám metod – metodám pevného rozkladu a hierarchickým metodám.

Obrázek č. 1: Klasifikační strom vytvořený metodou CHAID (Pearsonova statistika)



Zdroj: vlastní zpracování

Tabulka č. 5: Kontingenční tabulka pro vztah skupiny metod a výpočtu vzdálenosti pro metody pevného shlukování

		Skupina metod		Celkem
		metody pevného rozkladu	hierarchické metody	
Výpočet vzdálenosti	vzdálenosti objektů od centroidu nebo medoidu	4	0	4
	vzdálenosti mezi shluky	0	4	4
Celkem		4	4	8

Zdroj: vlastní zpracování

Obdobně se postupuje při aplikaci testu o nezávislosti s využitím věrohodnostního poměru, pouze jsou v klasifikačním stromu uváděny hodnoty této statistiky a odpovídající p-hodnoty, resp. p-hodnoty upravené Bonferroniho metodou.

V různých programových systémech se i při použití stejných algoritmů se stejným nastavením mohou výsledky lišit. Odlišné mohou být např. způsoby výběru vysvětlující proměnné v případě, kdy jsou získány dvě (příp. více) minimální p-hodnoty, jak je tomu v tabulce č. 2. V programovém systému IBM SPSS Decision Trees byla vybrána proměnná *shlukování*. Pokud bychom stejnou analýzu provedli v systému Statistica, byla by vybrána proměnná *vzdálenost*.

3.2 APLIKACE PRINCIPU ANALÝZY ROZPTYLU

Jak bylo zmíněno v části 2.2, při porovnání celkové a vnitroskupinové variability se v případě kategoriální vysvětlované proměnné využívá např. Giniho míra mutability. Na základě analýzy dat z tabulky č. 1 byl pomocí algoritmu CRT vytvořen klasifikační strom, který je uveden na obrázku č. 2. Ze znázorněného postupu je zřejmé, že pro klasifikaci byly využity dvě z původních tří proměnných. K rozlišení hierarchických metod a metod rozkladu byl využit způsob výpočtu vzdáleností. K rozlišení dvou skupin metod rozkladu byl použit typ shlukování.

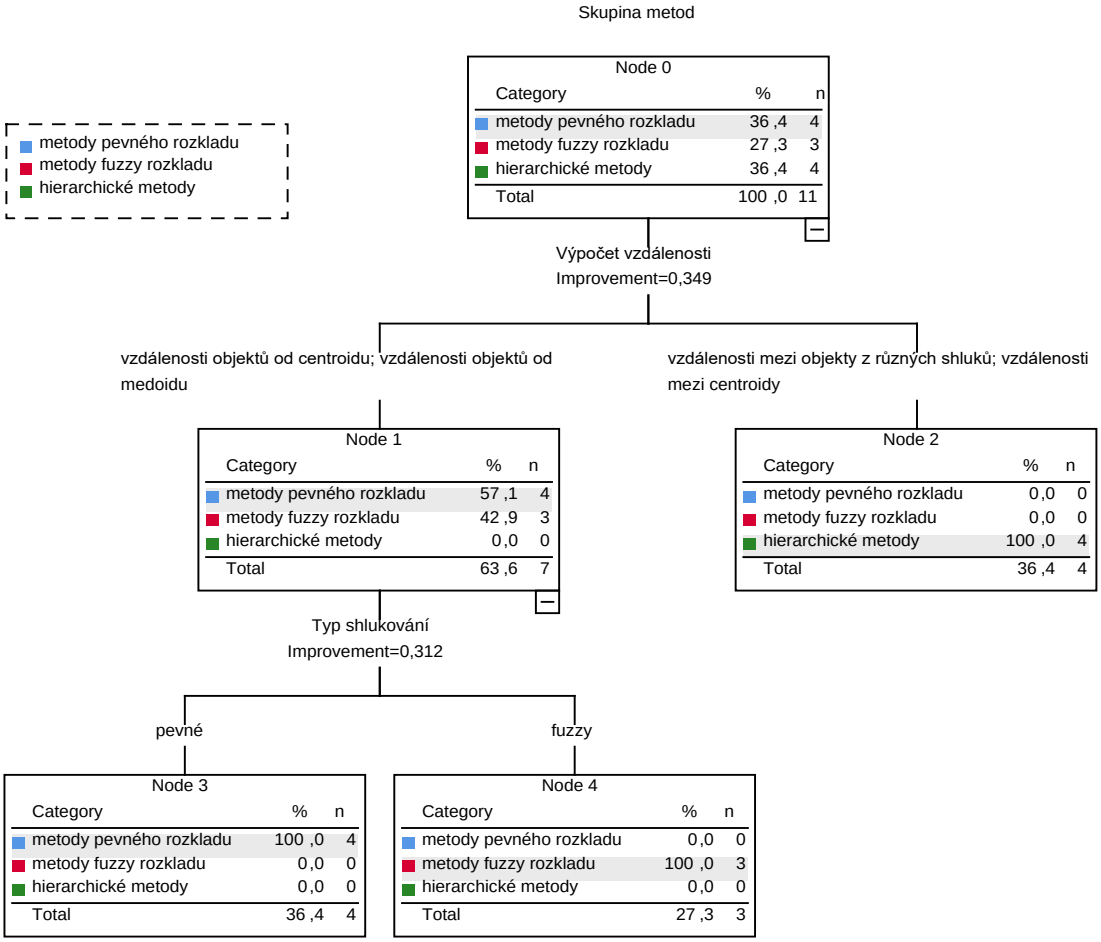
Hodnoty „improvement“ vyjadřují meziskupinovou variabilitu při rozdělení (větvení) určité skupiny objektů do menších skupin. Je vybráno takové rozdělení, kde je meziskupinová variabilita největší. V uzlu 0 jsou zahrnuty všechny objekty. Variabilita vysvětlované proměnné pomocí Giniho míry variability je 0,661 (viz tabulka č. 5). Tento uzel je rozdělen do uzlu 1 s 63,64 % objektů a s variabilitou 0,4898 a do uzlu 2 s 36,36 % objektů a s nulovou variabilitou. Průměrná variabilita ve skupinách je 0,6364·0,4898 + 0,3636·0 = 0,312. Rozdíl mezi celkovou variabilitou v původní množině objektů a vnitroskupinovou variabilitou při rozdělení do uzlů 1 a 2 (tj. meziskupinová variabilita) je tedy 0,661 – 0,312 = 0,349 (viz hodnota „improvement“ při rozdělení uzlu 0 do uzlů 1 a 2 na obrázku č. 2). Uzel 1 je rozdělen do dvou uzlů s nulovou variabilitou. Jako poslední hodnota je tedy uvedena variabilita v uzlu 1 (vážená), tj. 0,312 (viz hodnota „improvement“ při rozdělení uzlu 1 do uzlů 3 a 4 na obrázku č. 2). Pomocné výpočty variabilit pro uzly 0 a 1 jsou v tabulce č. 6 (podíl $\frac{n_{+ju}}{n_u}$ je označen jako p_{+ju} a platí, že $\frac{n_{+j1}}{n_1} = \frac{n_{1j0}}{n_{1+0}}$).

Tabulka č. 6: Pomocné výpočty pro klasifikační strom na obrázku č. 2

	Uzel 0		Uzel 1	
j	p_{+j0}	$p_{+j0}\ (1 - p_{+j0})$	p_{+j1}	$p_{+j1}\ (1 - p_{+j1})$
1	0,3636	0,2314	0,5714	0,2449
2	0,2727	0,1983	0,4286	0,2449
3	0,3636	0,2314	0,0000	0,0000
Součet	1,0000	0,6611	1,0000	0,4898

Zdroj: vlastní zpracování

Obrázek č. 2: Klasifikační strom vytvořený metodou CRT (Giniho míra)



Zdroj: vlastní zpracování

4. ZÁVĚR

Při výběru vysvětlujících proměnných v klasifikačních stromech by mělo být zohledněno, zda je vysvětlovaná proměnná nominální, nebo ordinální. V některých programových systémech je však zohlednění pouze částečné. Chí-kvadrát test se pro ordinální proměnnou buď nepoužívá vůbec (SAS Enterprise Miner), nebo se aplikuje pouze věrohodnostní poměr (IBM SPSS Decision Trees). V systému SAS Enterprise Miner jsou pro ordinální proměnnou výpočty Giniho míry a entropie speciálně upraveny. Problematika ordinální vysvětlované proměnné je postupně dále zkoumána a navržené postupy již byly implementovány v balíčcích v prostředí R.

V případě nominální vysvětlované proměnné nejsou dosud zohledněny všechny možnosti zkoumání závislostí. Používány jsou např. chí-kvadrát testy o nezávislosti, které hodnotí vzájemnou závislost proměnných. I když je jednostranná závislost její součástí, vhodnější by bylo aplikování speciálních měr pro jednostrannou závislost, resp. příslušné testy o nezávislosti. Na druhou stranu je třeba konstatovat, že tyto míry jsou založeny na zkoumání variability vysvětlované proměnné a jejího rozkladu, což je stejný princip jako při využití Giniho míry a entropie.

Dosud není speciálně řešena problematika diskrétní kvantitativní vysvětlované proměnné, pro kterou lze aplikovat neparametrické testy zaměřené na jednostrannou závislost, např. Kruskalův-Wallisův test. Při rozkladu variability je možné kromě rozptylu použít i jiné míry, např. Giniho průměrnou diferenci.

Poděkování

Tento článek byl připraven za podpory projektu IGA F4/44/2018 Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze.

LITERATURA

[1] AGRESTI, A.: Measures of nominal-ordinal association. In: Journal of the American Statistical Association, 1981, č. 375, s. 524 – 529.
[2] ARCHER, K. J.: rpartOrdinal: An R package for deriving a classification tree for predicting an ordinal response. In: Journal of Statistical Software, 2010, č. 7, s. 1 – 17. [online]. [cit. 30. 3. 2020]. Dostupné na: http://www.jstatsoft.org/v34/i07/.
[3] BLAIR, J. – LACY, M. G.: Statistics of ordinal variation. In: Sociological Methods & Research, 2000, č. 3, s. 251 – 280.
[4] GALIMBERTI, G. – SOFFRITTI, G. – MASO, M. D.: Classification trees for ordinal responses in R: The rpartScore package. In: Journal of Statistical Software, 2012, č. 10, s. 1 – 25. [online]. [cit. 30. 3. 2020]. Dostupné na: http://www.jstatsoft.org/v47/i10/.
[5] GINI, C. W.: Variability and Mutability. Contribution to the Study of Statistical Distributions and Relations. Studi Economico-Giuridici della R. Universita de Cagliari, 1912.
[6] JANITZA, S. – TUTZ, G. – BOULESTEIX, A.-L.: Random forest for ordinal responses: Prediction and variable selection. In: Computational Statistics & Data Analysis, April, 2016, s. 57 – 73.
[7] LABUDOVÁ, V.: Rozhodovacie stromy ako prediktívna modelovacia technika. In: Slovenská štatistika a demografia, 2017, č. 3, s. 60 – 76.
[8] LACY, M. G.: An explained variation measure for ordinal response models with comparisons to other ordinal R^2 measures. In: Sociological Methods & Research, 2006, č. 4, s. 469 – 520.
[9] LÖSTER, T.: Různé způsoby stanovení počtu shluků ve shlukové analýze. In: Slovenská štatistika a demografia, 2017, č. 3, s. 47 – 59.
[10] PICCARRETA, R.: A new measure of nominal-ordinal association. In: Journal of Applied Statistics, 2001, č. 1, s. 107 – 120.
[11] PICCARRETA, R.: Classification trees for ordinal variables. In: Computational Statistics, 2008, č. 3, s. 407 – 427.
[12] ŘEHÁK, J. – ŘEHÁKOVÁ, B.: Analýza kategorizovaných dat v sociologii. Praha: Academia, 1986. 397 s.
[13] ŘEZANKOVÁ, H.: Analýza dat z dotazníkových šetření. 4. vydání. Praha: Professional Publishing, 2017. 225 s. ISBN 978-80-906594-8-3.
[14] ŠULC, Z. – ŘEZANKOVÁ, H.: Comparison of similarity measures for categorical data in hierarchical clustering. In: Journal of Classification, 2019, č. 1, s. 58 – 72.

RESUMÉ

Článek pojednává o způsobech výběru vysvětlujících proměnných v klasifikačních stromech. Zaměřuje se jednak na dnes již poměrně dobře známé přístupy, kdy se předpokládá nominální vysvětlovaná proměnná. Vysvětlující proměnné jsou vybírány buď na základě chí-kvadrát testu (v programových systémech jsou obvykle nabízeny testy s využitím Pearsonovy statistiky a věrohodnostního poměru), nebo pomocí rozkladu variability, obdobně jako je zkoumán rozklad rozptylu v případě kvantitativní proměnné. Protože je však vysvětlovaná proměnná kategoriální, využívá se buď Giniho míra mutability, nebo entropie. Tyto známé přístupy jsou ilustrovány na analýze jednoduchého datového souboru v programovém systému IBM SPSS Decision Trees pomocí algoritmů CHAID a CRT. Kromě toho článek poukazuje také na současné trendy ve výzkumu v oblasti klasifikačních stromů a náhodných lesů, kterými jsou aplikace speciálních přístupů pro ordinální vysvětlovanou proměnnou.

RESUME

The paper deals with ways of explanatory variable selection in classification trees. It focuses on the well-known approaches when the nominal explanatory variable is expected. Explanatory variables are selected either on the basis of a chi-square test (in software systems the Pearson statistics and the likelihood ratio are usually available) or by means of variability decomposition, similarly as the variance decomposition is investigated in the case of the quantitative variable. However, for the reason that the target variable is categorical, either the Gini measure of mutability or the entropy are applied. These well-known approaches are illustrated on the analysis of the simple dataset using the CHAID and CRT algorithms in the IBM SPSS Decision Trees software system. Moreover, the actual research trends in the field of classification trees and the random forests are the applications of special techniques for the ordinal target variable.

PROFESNÍ ŽIVOTOPIS

Prof. Ing. Hana Řezanková, CSc., *absolvovala obor ekonomicko-matematické výpočty na Vysoké škole ekonomické v Praze, kde pôsobí v súčasnej dobe na katedře statistiky a pravděpodobnosti Fakulty informatiky a statistiky. Je členkou vedeckej rady a akademického senátu na tejto fakulte a predsedníčkou odborovej rady pre doktorský študijný program statistika. V rokoch 2013 – 2017 bola predsedníčkou České statistické společnosti a v rokoch 2015 – 2019 členkou České statistické rady. Vo svojej vedecko-výskumnej činnosti sa zameriava na analýzu kategoriálnych údajov a na metódy zhlukovej analýzy. Je autorkou či spoluautorkou niekoľkých knižných publikácií.*

KONTAKT

hana.rezankova@vse.cz