Získávání znalostí pomocí dvoukrokové klasifikace

Marina Stecenková

Katedra statistiky a pravděpodobnosti, FIS,
Vysoká škola ekonomická v Praze, nám. W. Churchilla 4, 130 67, Praha
Marina.Stecenkova@vse.cz

Abstrakt. Cílem příspěvku je představení metodologie, jenž kombinuje výstupy analýzy latentních tříd a klasifikačního stromu CHAID, a to za účelem zlepšení porozumění analyzovaných dat. Podstata spočívá v tom, že nejprve je množina objektů pomocí analýzy latentních tříd rozdělena do shluků objektů s podobnými vlastnostmi. Následně pro lepší profilaci vytvořených skupin je použita klasifikační metoda CHAID, kde vysvětlovanou proměnnou je právě vytvořená latentní proměnná, tj. proměnná určující příslušnost každého objektu do jednoho z vytvořených shluků. Pro názornou ilustraci této metodologie byla použita data z *Výběrového šetření pracovních sil* 2011.

Klíčová slova: klasifikace, analýza latentních tříd (LCA), algoritmus CHAID

1 Úvod

Kategorizace klientů a následný popis charakteristických rysů vytvořených skupin je v dnešní době nedílnou součástí ekonomické (obchodní) strategie řady velkých společností a další malé a střední podniky se k danému trendu postupně připojují. Poznání svých klientů je jednou z klíčových otázek pro dosažení úspěchu. S tím souvisí i rozvoj analytických nástrojů a v neposlední řadě i hledání nových postupů.
Cílem příspěvku je představit metodologii, která kombinuje vlastnosti analýzy latentních tříd (LCA, latent class analysis) a algoritmu CHAID, kde výstup z LCA je vstupem pro CHAID analýzu. Sledované objekty jsou tedy analyzovány ve dvou krocích. V prvním kroku jsou vytvořeny shluky (LCA) a ve druhém kroku jsou tyto shluky profilovány (CHAID ale i samotné LCA). Výhodou kombinace těchto algoritmů je získání dobře interpretovatelných výstupů, a to pravděpodobnostního profilu jednotlivých segmentů díky analýze LCA a klasifikačního stromu CHAID, který reprodukuje vytvořené pravděpodobnostní segmenty. Pro ilustraci jsou použita data z *Výběrového šetření pracovních sil* prováděného v ČR v roce 2011.

2 Dvoukroková klasifikace

Přístup dvoukrokové klasifikace je založen na analýze dat s využitím dvou různých typů klasifikačních metod. Nejprve se objekty (respondenti, zákazníci) z dané datové množiny pomocí LCA rozdělí podle vybraných faktorů do vhodného počtu shluků tak, aby si objekty spadající do stejného shluku byly co nejvíce podobné z hlediska

hodnot vybraných faktorů a zároveň aby rozdíly mezi shluky byly co největší. Vytvořené shluky je potřeba následně popsat (profilovat). Kromě klasického specifikování shluků na základě podmíněných pravděpodobností, které jsou výstupem LCA, nabízí Magidson a Vermunt [2] využít k popisu algoritmus CHAID, a to tak, že by naopak podle hodnot stejných faktorů byla odhadována příslušnost objektů k vytvořeným shlukům. Použití algoritmu CHAID umožňuje profilaci shluků i pomocí jiných faktorů, než které byly použity pro vytvoření shluků.

2.1 Analýza latentních tříd (LCA, latent class analysis)

Hlavní myšlenkou analýzy latentních tříd je, že vzájemná závislost mezi pozorovanými faktory je zapříčiněna vlivem skrytých (neměřitelných, latentních) faktorů. Jinými slovy se vychází z předpokladu, že objekty analyzovaného souboru pochází z jistého počtu homogenních skupin, které ovšem nejsou dopředu pozorovatelné. Cílem LCA je identifikace těchto skupin, u kterých se předpokládají odlišná pravděpodobnostní rozdělení. Podrobněji např. v [1] nebo [2].

2.2 Algoritmus CHAID

Algoritmus CHAID (Chi-squared automatic interaction detection) je nebinární klasifikační strom, který využívá χ^2 statistiku ke štěpení množiny objektů analyzovaného souboru od kořene stromu přes nekoncové uzly až ke koncovým uzlům. χ^2 statistika umožňuje měřit homogenitu v uzlech a rozdělit data na další podmnožiny. Cílem je, aby koncové uzly byly tvořeny alespoň z větší části objekty ze stejných, předem známých, tříd. Více např. v [3].

2.3 Segmentace nezaměstnaných podle způsobu hledání zaměstnání

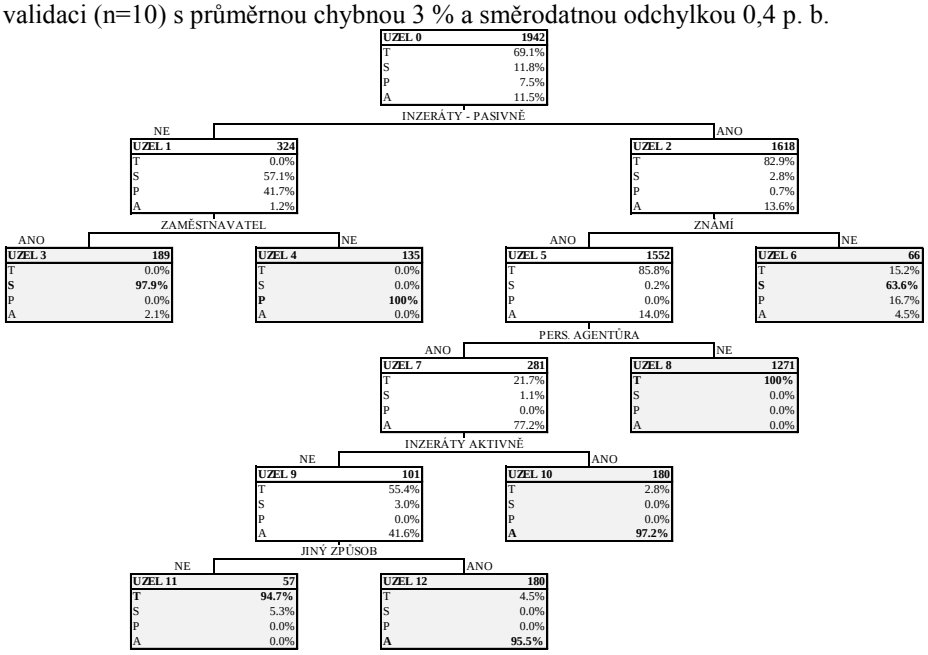
Výše popsaná metodologie byla aplikována na datech získaných z *Výběrového šetření pracovních sil* 2011. Cílem bylo prozkoumat způsob hledání nového zaměstnání osob bez práce, a to s předpokladem, že nezaměstnané lze rozdělit do několika skupin s podobným vzorcem chování. LCA byla provedena na sedmi binárních faktorech odpovídajících způsobu hledání nového zaměstnání nezaměstnaných osob. Celkem 1942 osob odpovídalo, zda hledaly své budoucí zaměstnání pomocí registrace na úřadu práce, přes personální agentury, získáváním kontaktů od známých, přímým kontaktem potenciálního zaměstnavatele, aktivním vypisováním a odpovídáním na inzeráty s nabídkou práce, pouze pasivním procházením nabídek práce nebo jiným způsobem.
Z hlediska věcné logiky a na základě nejnižších hodnot Akaikeho a Bayesova informačního kritéria byl zvolen model se čtyřmi latentními třídami (v systému LatentGold 4.0). Tab. č. 1 zobrazuje pravděpodobnostní profily jednotlivých tříd, které byly pojmenovány jako *Typický* (do této skupiny bylo zařazeno 69 % nezaměstnaných), *Specifický* (12 %), *Pasivní* (7 %) a *Aktivní* (12 %). Prevalence určuje pravděpodobnost, že nezaměstnaný bude patřit do dané skupiny. Kdežto

podmíněné pravděpodobnosti udávají, jaká je pravděpodobnost, že nezaměstnaný zvolí daný způsob hledání zaměstnání za předpokladu, že patří do určité skupiny.

Tabulka 1. Podíly nezaměstnaných, kteří hledají daným způsobem nové zaměstnání, a pravděpodobnostní profily jednotlivých segmentů.

Způsob hledání práce	Podíl nezaměstnaných	Latentní třída			
		Typický	Specifický	Pasivní	Aktivní
Úřad práce	0.90	0.92	0.82	0.85	0.96
Pers. Agentury	0.17	0.06	0.19	0.03	0.60
Zaměstnavatel	0.81	0.81	0.98	0.02	0.96
Známí	0.91	0.98	0.80	0.40	0.98
Inzeráty – aktivně	0.35	0.35	0.02	0.01	0.80
Inzeráty – pasivně	0.83	1.00	0.42	0.11	0.99
Jiný	0.37	0.33	0.26	0.14	0.70
Prevalence		0.58	0.16	0.08	0.18

V dalším kroku byl v systému SPSS 16.0 konstruován klasifikační strom CHAID, který zařadil správně 98 % nezaměstnaných do získaných latentních tříd při křížové validaci (n=10) s průměrnou chybnou 3 % a směrodatnou odchylkou 0,4 p. b.



Obrázek 1. Klasifikace do čtyř latentních tříd (výstup LCA) podle faktorů odpovídajících způsobu hledání nového zaměstnání (stejné jako u LCA).

Na základě pravděpodobnostních profilů a větvení klasifikačního stromu byly čtyři skupiny profilovány následovně. Pro skupinu *Typických*, do které patří téměř 70 % nezaměstnaných, je charakteristické pasivní prohledávání inzerátů, poptávání se lidí ve svém okolí, přímé oslovení zaměstnavatele a zároveň nevyužívání personálních agentur. Skupina *Specifických* nesleduje nabídky práce a nejčastěji přímo kontaktuje zaměstnavatele nebo se poptává u známých. Nejvýraznější rozdíly jsou vidět mezi skupinami *Aktivní* a *Pasivní*. Zatímco *Aktivní* využívají k hledání nového zaměstnání všech dostupných prostředků, tak *Pasivní* jsou pouze zaregistrování na úřadu práce. Použitím algoritmu CHAID mohou být jednotlivé skupiny dále profilovány z hlediska jiných faktorů (např. demografických, socio-ekonomických, atd.).

3 Závěr

Správné určení počtu shluků a jejich následná profilace je vždy obtížný úkol. Kritéria určující optimální počet shluků nám sice pomáhají z daných modelů vybrat ten lepší, ale může se také stát, že mezi srovnávanými modely nebude ani jeden dobrý z hlediska věcné interpretace. Metodologie popsaná v tomto příspěvku přináší hlavní výhodu právě v urychlení a zlepšení profilace získaných segmentů.

Poděkování

Tato práce vznikla s podporou grantu č. IGA F4/104 /2014.

English summary

Obtaining knowledge using two step classification

The aim of the paper is to introduce using the results of latent class analysis (LCA) and the CHAID algorithm for segmentation and profiling of researched objects where the output of LCA (classes) serves as an input for CHAID. Thus data will be segmented in two steps where for each step the same variables will be used. To illustrate this methodology, data from the periodical Labour Force Survey held in spring 2011 in the Czech Republic are used.

Reference

1. Collins L. M., Lanza S. T. *Latent Class and Latent Transition Analysis for the Social, Behavioral, and Health Sciences*. Wiley, New York, 2010, ISBN: 978-0-470-22839-5, str. 39-44.
2. Magidson J., Vermunt J. K. *Latent Class Analysis*, The Sage Encyclopedia of Social Science Research Methods, NewBury Park: Sage Publications, Inc., 2004.
3. Tufféry S. *Data Mining and Statistics for Decision Making*. Wiley, United Kingdom, 2011. ISBN 978-0-470-68829-8, str. 313-328.