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.

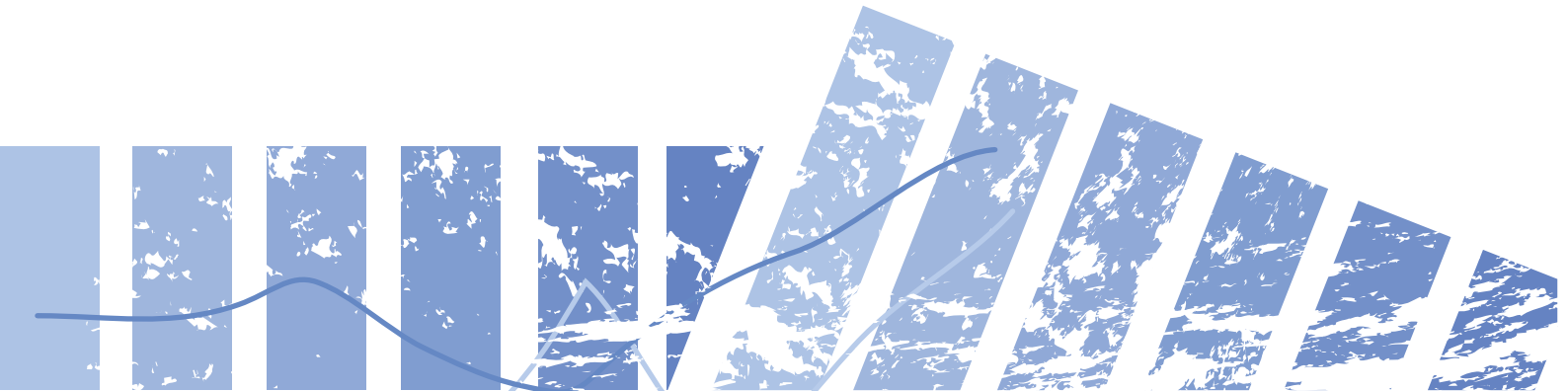
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 5

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

Strany/Pages: 47 – 59

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



Tomáš LÖSTER
Vysoká škola ekonomická v Praze

RŮZNÉ ZPŮSOBY STANOVENÍ POČTU SHLUKŮ VE SHLUKOVÉ ANALÝZE

VARIOUS METHODS OF DETERMINING THE NUMBER OF CLUSTERS IN CLUSTER ANALYSIS

ABSTRAKT
V současné odborné literatuře existuje celá řada způsobů, jak stanovit optimální počet shluků. Běžný způsob, který je často v literatuře uváděn a v praxi využíván, spočívá v nalezení počtu shluků na základě grafu – dendrogramu. Jedná se však o značně subjektivní záležitost, a tak bývá doporučeno využít některý z koeficientů pro stanovení počtu shluků. Těchto koeficientů je celá řada a neexistuje jednoznačné pravidlo, které by definovalo použitelnost daných koeficientů. Cílem tohoto článku je ukázat vybrané možnosti stanovení počtu shluků v různých podmínkách při pevném hierarchickém shlukování.

ABSTRACT
In the current specialised literature, there are many methods to determine the optimal number of clusters. A common method which is often mentioned in literature and used in practice lies in determining the number of clusters on the basis of the graph - the dendrogram. However, it is a quite subjective matter, thus it is recommended to use one of the coefficients for determining the number of clusters. There is a large number of such coefficients and there is no clear rule for the use of these coefficients. The aim of this article is to show the selected possibilities of determining the number of clusters under various conditions in a stable hierarchical clustering.

KLÍČOVÁ SLOVA
shlukování, hodnocení shlukování, koeficienty pro stanovení optimálního počtu shluků

KEY WORDS
Clustering, Evaluation of clustering, Coefficients for determining the optimal number of clusters.

1. ÚVOD

Základní úlohou, kterou řeší mnoho vědních disciplín, je vytváření skupin objektů. Ty mohou být representovány zákazníky, pacienty, automobily, dokumenty, atd. K vytváření skupin mohou být využity různé matematicko-statistické metody a postupy. Pokud je objekt (pozorování) zařazován do existující skupiny, využívá se k tomu diskriminační analýza, pokud je objekt zařazován do tříd, které nemusí být předem známé, využívá se k tomu shluková analýza. Ta představuje vícerozměrnou statistickou metodu, jejímž cílem je vytváření skupin objektů, které se nazývají shluky. Uplatnění shlukové analýzy lze najít v mnoha odvětvích. Často se využívá při řešení ekonomických úloh. Vytváří se skupiny klientů podle různé strategie či rizika, shlukují se firmy, země, atd.

Základním cílem metod shlukové analýzy je vytvářet skupiny objektů (shluky), které jsou charakterizovány pomocí různých proměnných. Při vytváření shluků je důležité, aby si objekty, které jsou zařazeny uvnitř jednoho shluku, byly co nejvíce podobné a objekty, které jsou zařazeny do dvou různých shluků, si byly co nejméně podobné.

V současné odborné literatuře existuje mnoho metod shlukové analýzy. Ty mohou být členěny pomocí různých kritérií. Vývoj metod shlukové analýzy je spojen jednak se vznikem nových metod, jednak s modifikacemi stávajících metod. V současné době tak díky celé řadě softwarových produktů existuje velké množství metod a postupů, které může konečný uživatel aplikovat. Neexistuje však pravidlo, které by určilo, jak zvolit vhodnou kombinaci metod a algoritmů k tomu, aby výsledné rozdělení objektů do shluků bylo nejlepší. Počet výsledných skupin objektů často není předem známý, a proto součástí shlukové analýzy bývá stanovení optimálního počtu shluků, do kterého mají být objekty klasifikovány. Cílem tohoto článku je ukázat možnosti stanovení počtu shluků na základě vybraných koeficientů, které jsou dostupné v softwaru a tím pádem využívané z řad analytiků.

2. TEORETICKÝ RÁMEC

Shluková analýza je oblíbená vícerozměrná metoda, která se využívá v řadě ekonomických oblastí, mezi které může být zařazena například klasifikace domácností podle typu a vztahu k materiální deprivaci, viz například [17]. Samotný proces shlukování a jeho výsledky jsou velmi intenzivně závislé na volbě proměnných, pomocí kterých jsou jednotlivé objekty charakterizovány. Jak je uvedeno v [16], výsledkem shlukové analýzy není stanovení významných či nevýznamných proměnných, nýbrž vytvoření shluků, na základě vhodně vybraných vlastností objektů. Základní členění tradičních metod shlukování spočívá v rozdělení na *hierarchické* (slouží k vytváření stromovité struktury) a *nehierarchické* metody shlukování, které představují metody rozkladu, viz například [12], [16]. Speciálním případem je tzv. *fuzzy* shlukování, kde zařazení objektu do shluku je dáno tzv. mírou příslušnosti. Ta představuje hodnotu z intervalu od 0 do 1, která vyjadřuje příslušnost, že daný objekt je zařazen do daného shluku. Vyplývá z toho, že objekt může být zařazen do více shluků současně a míra příslušnosti představuje „pravděpodobnost“, že objekt bude klasifikován do daného shluku.

Pro shlukovou analýzu představují klíčovou informaci *míry podobnosti*. Při měření podobností záleží na typu proměnných, které charakterizují jednotlivé objekty. Mezi nejznámější charakteristiky lze zařadit Euklidovu vzdálenost či Mahalanobisovu vzdálenost, které se zejména používají v případě, že jsou objekty charakterizovány pomocí kvantitativních proměnných. Euklidova vzdálenost není vhodná pro případ, kdy jsou jednotlivé proměnné, které charakterizují jednotlivé objekty, velmi silně korelované. Jak je uvedeno například v [4], Mahalanobisova vzdálenost, na rozdíl od Euklidovy míry vzdálenosti, odstraňuje problém, který vzniká při použití nestandardizovaných dat, které mohou způsobit rozdíly mezi shluky, v důsledku odlišností měrných jednotek. Tato míra vzdálenosti je navíc použitelná i tehdy, jestliže jsou jednotlivé proměnné vzájemně závislé.

Jak je uvedeno například v [16] nebo [12], mezi nejznámější a nejpoužívanější metody shlukování lze zařadit například metodu nejbližšího souseda, metodu nejvzdálenějšího souseda, metodu průměrné vzdálenosti, Wardovu metodu, atd.

Tyto metody se liší nejen dobou vzniku, ale také přístupem ke shlukování. Jejich podrobný popis lze nalézt například v [4], [16].

3. KOEFICIENTY PRO STANOVENÍ OPTIMÁLNÍHO POČTU SHLUKŮ PŘI HIERARCHICKÉ SHLUKOVÉ ANALÝZE

Součástí shlukové analýzy velmi často bývá stanovení počtu shluků, do kterých mají být dané objekty rozděleny. Ke stanovení optimálního počtu shluků existuje celá řada kritérií a postupů. Použití různých metod shlukování může přinášet rozdílná rozdělení objektů do shluků. Problematice stanovení optimálního počtu shluků pro případ, že jsou shluky charakterizovány kvantitativními proměnnými, je věnována celá řada odborných článků, mezi které patří například [1], [2], [5], [6], [7], [8], [14], atd. Problematice stanovení optimálního počtu shluků v případě jiných proměnných se věnují například v [9] či [10]. V současné odborné literatuře neexistuje jednoznačné pravidlo, které by určilo použití konkrétních koeficientů v různých podmínkách. Různí autoři vytváří či modifikují koeficienty, někdy částečně své a vybrané koeficienty porovnávají s již existujícími. Mezi vybraná kritéria pro stanovení optimálního počtu shluků je možné zařadit Daviesův-Bouldinův (DB) index, Dunnův index, RMSSTD index, CHF index, PTS index.

V této části článku jsou dále popsány výše uvedené vybrané koeficienty pro disjunktní shlukování. Předpokládá se rozdělení množiny *n* objektů do *k* disjunktních shluků, přičemž každý objekt je zařazen právě do jednoho shluku. U koeficientů pro stanovení počtu shluků většinou nezáleží, zda jsou shluky výsledkem metod rozkladu anebo výsledkem hierarchického shlukování, viz [12]. Některé koeficienty jsou však určeny výlučně pro shluky, které vznikly na základě hierarchického shlukování.

Použití **Daviesova-Bouldinova indexu**, jak je uvedeno v [11], nezávisí na vybrané metodě shlukování. Aby bylo možné určit hodnoty Daviesova-Bouldinova indexu, je nejprve nutné definovat tzv. *disperzi h*-tého shluku S_h, která se stanoví jako

$$S_h = \sqrt{\frac{\sum_{x_i \in C_h} D^2(\mathbf{x}_i, \bar{\mathbf{x}}_h)}{n_h}},$$

kde význam jednotlivých symbolů je následující: n_h je počet objektů v *h*-tém shluku; $\mathbf{x}_i \in \mathrm{C}_h$ představuje označení, že *i*-tý objekt se nachází v shluku C_h, $\bar{\mathbf{x}}_h$ je centroid *h*-tého shluku, a pro kterou platí následující podmínky

$S_h \geq 0$,

$S_h = 0$, v případě, že jsou objekty ve shluku charakterizovány identickými vlastnostmi.

Pro měření vzdáleností shluků *D* platí

$$D_{hh'} = D(\bar{\mathbf{x}}_h, \bar{\mathbf{x}}_{h'}),$$

$$D_{hg} = D(\bar{\mathbf{x}}_h, \bar{\mathbf{x}}_g),$$

kde $\bar{\mathbf{x}}_h$ je centroid *h*-tého shluku, $\bar{\mathbf{x}}_{h'}$ je centroid *h´*-tého shluku a $\bar{\mathbf{x}}_g$ je centroid *g*-tého shluku.

Nechť **míra podobnosti** mezi *h*-tým a *h´*-tým shlukem se značí $A_{hh´}$ a je založena na disperzích těchto shluků a podle [12] musí splňovat následující podmínky:

1. $A_{hh´} \geq 0$,
2. $A_{hh´} = A_{h´h}$,
3. $A_{hh´} = 0$, pokud $S_h = S_{h´}$,
4. $A_{hh´} > A_{hg}$, pokud $S_{h´} = S_g$ a $D_{hh´} < D_{hg}$,
5. $A_{hh´} > A_{hg}$, pokud $S_{h´} > S_g$ a $D_{hh´} = D_{hg}$,

kde S_h, $S_{h´}$, S_g jsou disperze *h*-tého, *h´*-tého a *g*-tého shluku, $D_{hh´}$, D_{hg} jsou vzdálenosti mezi jednotlivými shluky.

V následujícím kroku se určí míra podobnosti mezi *h*-tým a *h´*-tým shlukem podle vzorce

$$A_{hh´} = \frac{S_h + S_{h´}}{D_{hh´}} .$$

Maximální míra podobnosti mezi shluky *h* a *h´* se dále označí jako A_h, tj.

$$A_h = \max_{h, h \neq h´} A_{hh´} .$$

Daviesův-Bouldinův index se nakonec určí jako aritmetický průměr maximálních měr podobností A_h, tedy podle vzorce

$$I_{\mathrm{DB}}(k) = \frac{\sum_{h=1}^{k} A_h}{k} ,$$

kde *k* je počet shluků.

Vyhodnocení optimálního počtu shluků se pomocí Daviesova-Bouldinova indexu provádí nalezením minimální hodnoty tohoto indexu, která indikuje kompaktní a dobře separované shluky. Základem je stanovit hodnoty tohoto indexu na předem stanoveném maximálním počtu shluků, tj.

$$I_{\mathrm{DB}}(k^*) = \min_{2 \leq k \leq n-1} I_{\mathrm{DB}}(k) ,$$

kde *k** je optimální počet shluků.

Další možností, jak stanovit počet shluků, je využít **Dunnův index**. Pomocí tohoto indexu je opět možné najít kompaktní a dobře separované shluky, viz [3].

Vzdálenost mezi *h*-tým a *h´*-tým shlukem je definována jako minimální vzdálenost dvou objektů z těchto různých shluků

$$D_{hh´} = \min_{\mathbf{x}_i \in C_h, \mathbf{x}_j \in C_{h´}} D(\mathbf{x}_i, \mathbf{x}_j) .$$

Nechť M_h je definována jako maximální vzdálenost dvou objektů ze stejného (h-tého) shluku, tedy

$$M_h = \max_{\mathbf{x}_i, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j) .$$

Dunnův index je následně definován jako

$$I_D(k) = \min_{1 \le h \le k} \left\{ \min_{1 \le h' \le k} \frac{D_{hh'}}{\max_{1 \le h \le k} M_h} \right\} .$$

Při stanovení optimálního počtu shluků se, na rozdíl od Daviesova-Bouldinova indexu, hledá maximální hodnota tohoto indexu v rámci předem stanoveného počtu shluků, který je opět menší než počet objektů, tj.

$$I_D(k^*) = \max_{2 \le k \le n-1} I_D(k) .$$

Vysoké hodnoty tohoto indexu indikují kompaktní a dobře separované shluky.

V situaci, kdy jsou jednotlivé objekty charakterizovány pouze pomocí t kvantitativních proměnných, pro měření variability je možné využít rozptyl. Další koeficienty pro hodnocení výsledků shlukování jsou založeny na rozkladu celkové variability na vnitroshlukovou a mezishlukovou složku variability. Jde o analogii analýzy rozptylu. Do této skupiny patří například **RS index** (též R-kvadrát, RSQ index), který je možné využít pro srovnávání různých postupů shlukování, nejen při hierarchickém shlukování. Tento index je možné využít také k vyjádření kvality shluků, viz [12]. Jeho myšlenka je založena na rozkladu celkového součtu čtverců na vnitroshlukovou a mezishlukovou složku variability.

Nechť jsou označeny následující součty čtverců, viz [12]:
SS_B = součet čtverců mezi shluky (charakteristika mezishlukové variability),
SS_W = součet čtverců uvnitř shluků (charakteristika vnitroshlukové variability),
SS_T = celkový součet čtverců (charakteristika celkové variability).

Jednotlivé součty čtverců se stanoví podle následujících vzorců

$$SS_W = \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} \sum_{l=1}^{t} (x_{it} - \bar{x}_{hl})^2 ,$$

$$SS_T = \sum_{i=1}^{n} \sum_{t=1}^{m} (x_{it} - \bar{x}_t)^2 ,$$

$$SS_B = SS_T - SS_W .$$

RS index je následně definován jako podíl mezishlukového a celkového součtu čtverců, tj. podle vzorce

$$I_{RS} = \frac{SS_B}{SS_T} = \frac{SS_T - SS_W}{SS_T} .$$

Tento koeficient není vhodné používat pro stanovení optimálního počtu shluků, protože s rostoucím počtem shluků (tj. se snižujícím se počtem objektů v nich) se dílčí shluky stávají více homogenní. V důsledku toho dochází k nárůstu mezishlukového součtu čtverců, a tedy k nárůstu hodnoty tohoto indexu. Z tohoto

důvodu je vhodné používat tento index pro srovnání úspěšnosti různých shlukovacích metod. Index nabývá hodnot z intervalu od 0 do 1, přičemž hodnota 0 vyjadřuje, že nejsou žádné rozdíly mezi shluky, hodnota 1 vyjadřuje významný rozdíl mezi shluky, které jsou homogenní.

RMSSTD index (*root-mean-square standard deviation index*) je index, který opět využívá rozklad celkového součtu čtverců na dílčí složky. Měří homogenitu nových shluků a jeho výpočet je založen pouze na vnitroshlukové variabilitě, viz [5]. Jeho výpočet je definován podle vzorce

$$I_{\mathrm{RMSSTD}}\ (k) = \sqrt{\frac{SS_{\mathrm{W}}}{t \cdot (n - k)}}\ .$$

Hodnoty tohoto koeficientu je možné využít také ke stanovení optimálního počtu shluků. Nízké hodnoty RMSSTD indexu opět indikují lepší rozdělení objektů do výsledných shluků. V případě, že tento index nabývá vysokých hodnot, jedná se o nehomogenní shluky. Při grafickém vyhodnocení hodnot tohoto indexu pro jednotlivé počty shluků se optimální počet shluků stanoví podle „bodu zlomu“ křivky.

Jak uvádí samotní autoři koeficientů v [6], při vyhodnocení výsledků shlukování pomocí těchto koeficientů je vhodné stanovit hodnoty všech těchto koeficientů současně, aby výsledné hodnocení byly co „nejobjektivnější“.

Další koeficient, který vychází z rozkladu celkového součtu čtverců, je **CHF index** (též pseudo *F* index), který byl navržen autory Calinski a Habarasz, viz [12]. Dále pak byl zkoumán autory Maulik a Bandyopadhyay, viz [14]. CHF index je definován jako podíl průměrné mezishlukové a průměrné vnitroshlukové variability, tj. podle vzorce

$$I_{\mathrm{CHF}}\ (k) = \frac{\dfrac{SS_{\mathrm{B}}}{k - 1}}{\dfrac{SS_{\mathrm{W}}}{n - k}} = \frac{(n - k) \cdot SS_{\mathrm{B}}}{(k - 1) \cdot SS_{\mathrm{W}}}\ .$$

Tento koeficient představuje analogii *F*-testu, který se používá v analýze rozptylu. Je možné jej využít pro stanovení optimálního počtu shluků. Vysoké hodnoty tohoto koeficientu indikují dobře separované shluky, tj. při stanovení optimálního počtu shluků se hledá maximální hodnota tohoto indexu v rámci předem stanoveného počtu shluků

$$I_{\mathrm{CHF}}\ (k^*) = \max_{2 \le k \le n-1} I_{\mathrm{CHF}}\ (k)\ .$$

Tento koeficient byl také modifikován, viz [12], pro hodnocení výsledků shlukování v případě kvalitativních proměnných a proměnných různých typů. V těchto případech uvedený koeficient poskytoval nejlepší výsledky v porovnání se známým počtem shluků.

PTS index (též pseudo T-kvadrát index) opět využívá myšlenky rozkladu celkového součtu čtverců na jednotlivé složky. Je možné jej využít pro stanovení optimálního počtu shluků, viz [15]. Vychází z vyhodnocení spojení *h*-tého a *h´*-tého shluku. Stanoví se podle vzorce

$$I_{\mathrm{PTS}}(k) = \frac{SS_{\mathrm{B}_{hh'}}}{\frac{SS_{\mathrm{W}_h} + SS_{\mathrm{W}_{h'}}}{n_h + n_{h'} - 2}},$$

kde $SS_{\mathrm{B}_{hh'}}$ je mezishlukový součet čtverců a SS_{W_h} a $SS_{\mathrm{W}_{h'}}$ jsou vnitroshlukové součty čtverců.

Vyhodnocení se provádí tak, že v případě, že je pro *k* shluků hodnota tohoto indexu větší než pro (*k* – 1) a (*k* + 1) shluků zároveň, optimální počet shluků je (*k* + 1).

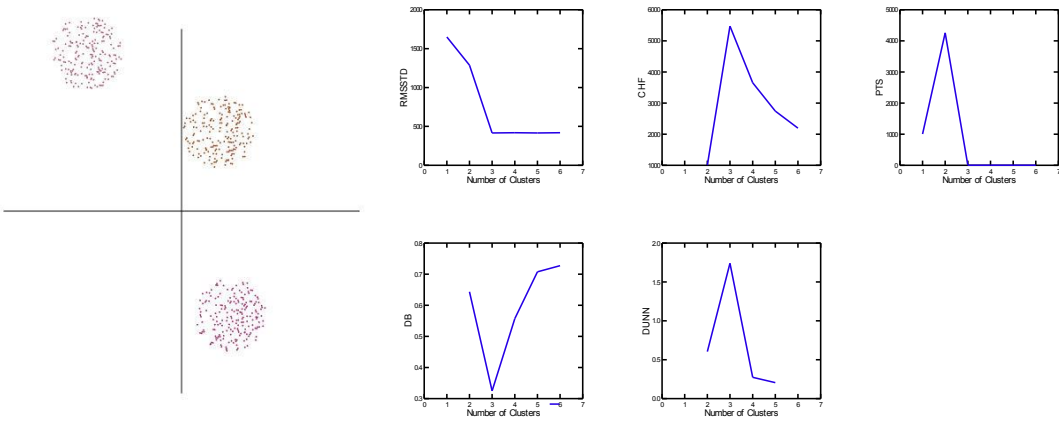
Ke shlukování lze využít celou řadu softwarových produktů. Mezi nejznámější a velmi často používané lze zařadit například systémy IBM SPSS, SAS, STATISTICA, S-PLUS, SYSTAT, STATGRAPHICS, atd. V nich jsou implementovány zejména tradiční metody shlukování, včetně případných metod rozkladu. Některé z nich, jako je třeba systém SAS, neumožňují uživateli volit příslušné kombinace shlukovacích metod a měr vzdálenosti. Systém nabízí kombinace jako dané.

4. STANOVENÍ POČTU SHLUKŮ V PRAKTICKÝCH ÚLOHÁCH

V této části článku je představen postup, jak vybrat počet shluků na základě výše popsaných koeficientů bez ohledu na zvolenou metodu shlukování a míru vzdálenosti. V každém z grafů bude v levé části zobrazeno reálné rozdělení objektů do shluků, které se liší barvou. V pravé části grafu je zobrazen grafický výstup pěti výše popsaných koeficientů ze systému SYSTAT pro dané rozdělení objektů do shluků. Není zde hodnocena úspěšnost vybraných metod shlukování v kombinaci s různými měrami vzdáleností, ale pouze schopnost stanovit počet shluků pro danou situaci.

První situace, která je zobrazena na obrázku č. 1, vyjadřuje tři dobře separované shluky. Z obrázku 1 vyplývá, že uvedené koeficienty jsou pro tři dobře separované shluky schopny správně odhalit jejich počet.

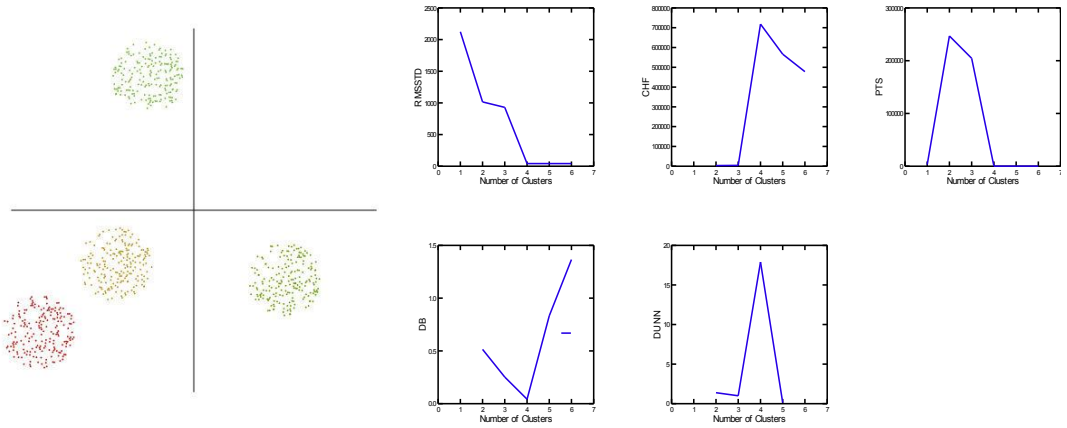
Obrázek č. 1: Tři dobře separované shluky



Zdroj: vlastní zpracování

Druhá situace, která je zobrazena na obrázku č. 2, zachycuje čtyři dobře separované shluky.

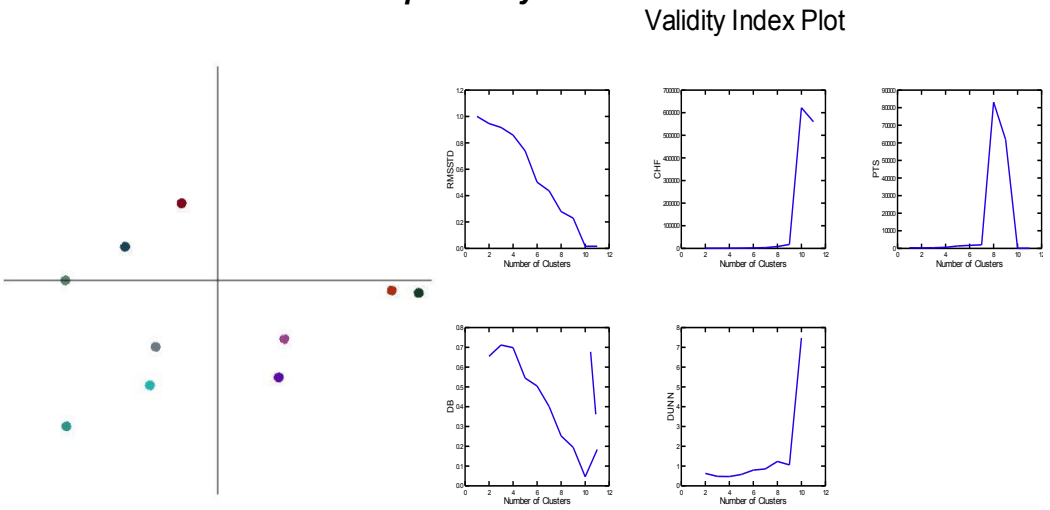
Obrázek č. 2: Čtyři dobře separované shluky



Zdroj: vlastní zpracování

I v této situaci je zřejmé, že uvedené koeficienty jsou schopny správně stanovit počet shluků. Další situace, která je zobrazena na obrázku č. 3, zachycuje deset dobře separovaných shluků. V případě deseti dobře separovaných shluků je opět zřejmé, že uvedené koeficienty jsou schopny správně odhalit počet shluků.

Obrázek č. 3: Deset dobře separovaných shluků

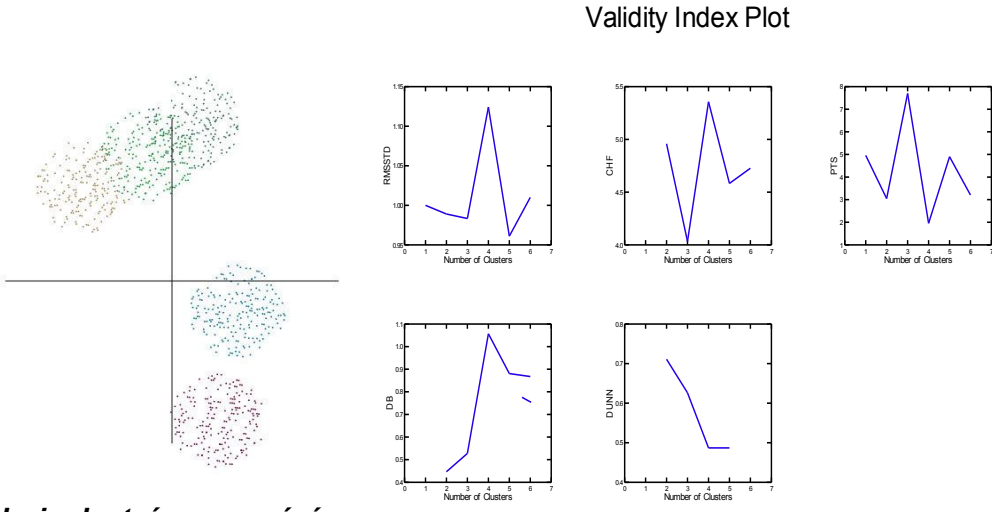


Zdroj: vlastní zpracování

Jak vyplývá z těchto příkladů, je možné konstatovat, že v případě dobře separovaných shluků se koeficienty chovají stabilně a počet shluků bývá dobře stanovitelný. Je zřejmé, že počet shluků je správně nalezen nejen v případě malého počtu výsledných shluků, ale i v případě většího počtu shluků.

Poslední popsaná situace, která je zobrazena na obrázku č. 4, zachycuje pět shluků, které se vzájemně překrývají. Prakticky to znamená, že se objekty různých barev (tedy z různých shluků) nachází v jedné oblasti (průnik dvou a více shluků).

Obrázek č. 4: Pět překrývajících se shluků



Zdroj: vlastní zpracování

V případě pěti překrývajících se shluků je již zřejmé, že uvedené koeficienty se jeví jako nestabilní, poskytují různé hodnoty a je tedy nutné hledat průnik co nejvyššího počtu koeficientů. V praktických úlohách se však může stát, že neexistuje žádný průnik a tak volba počtu shluků závisí na analytikovi.

Jak je z výše uvedených příkladů zřejmé, v případě dobře separovaných shluků koeficienty jejich počet odhalují správně. Úspěšnost koeficientů klesá v případě, že se výsledné shluky překrývají. Z tohoto důvodu je vhodné tuto situaci analyzovat podrobněji. Pro tuto analýzu byly vybrány soubory z databáze *The UCI Machine Learning Repository* (*http://archive.ics.uci.edu/ml/datasets.html*). Analýze byly podrobeny soubory, které jsou určené ke klasifikaci objektů s možností analyzovat všechny proměnné současně, tj. do výběru nevstupovaly soubory, které obsahovaly kvalitativní proměnné. Následnému hodnocení byly podrobeny ty soubory, u nichž je uveden skutečný počet výsledných shluků (zařazení objektů do shluků), aby bylo následně možné uskutečnit srovnání. V případě, že se v některých datových souborech vyskytly chybějící hodnoty u některých proměnných, uvedené objekty byly z dalších analýz vyřazeny. V případě nestejných měrných jednotek byla provedena standardizace pomocí tzv. Z-skórů. Pro účely vyhodnocení postupů pro stanovení optimálního počtu shluků byly vybrány a hodnoceny následující datové soubory, u nichž je uváděno zařazení objektů do shluků. Jedná se o soubory (seřazeny abecedně): *Abalone, Banknote Authentication, Blood Transfusion Service Center, Cardiotocography, Connectionist Bench (Vowel Recognition – Deterding Data), Energy Efficiency, Indian Liver Patient, Ionosphere, Iris, Musk (Version 1), QSAR Biodegradation, Statlog (Vehicle Silhouettes) a+b, Susy, Vertebral Column 2c, Vertebral Column 3c, Wall-Following Robot Navigation Data.* Uvedené soubory se týkají různých oblastí. Jedná se například o bankovky, pacienty, usně, atd. Při analýze každého souboru byly aplikovány následující metody: metoda nejbližšího souseda, nejvzdálenějšího souseda, metoda průměrné vazby, centroidní metoda a Wardova metoda. Každá z uvedených metod byla aplikována společně s Euklidovou vzdáleností a Mahalanobisovou vzdáleností, o které se uvádí, že odstraňuje potenciální problém s vzájemnou závislostí mezi proměnnými, které charakterizují jednotlivé objekty. V daných souborech se běžně vyskytuje překrytí shluků, tj. jeden objekt se nachází současně v prostoru dvou nebo více shluků.

V následujících tabulkách jsou uvedeny hodnoty úspěšností jednotlivých koeficientů, které byly získány srovnáním známého (skutečného) počtu shluků a nalezeného počtu shluků s danou kombinací metody a vzdálenosti pro všechny výše popsané soubory dohromady.

Tabulka č. 1: Úspěšnost koeficientů při použití Euklidovy vzdálenosti (v %)

Metoda/koeficient	RMSSTD	CHF	PTS	D-B	Dunn
Nejbližšího souseda	10,53	47,37	42,11	36,84	57,89
Nejvzdálenějšího souseda	26,32	15,79	21,05	63,16	42,11
Centroidní metoda	31,58	36,84	26,32	42,11	47,37
Průměrná vzdálenost	26,32	31,58	26,32	42,11	42,11
Wardova metoda	21,05	36,84	36,84	15,79	42,11

Zdroj: vlastní zpracování

Jak vyplývá z tabulky 1, při užití Euklidovy míry vzdálenosti v případě, že se shluky překrývají, je úspěšnost vybraných koeficientů nižší, než v případě dobře separovaných shluků. Nejlepších výsledků (63, 16 %) bylo dosaženo při použití Daviesova-Bouldinova indexu za současné aplikace s metodou nejvzdálenějšího souseda.

Tabulka č. 2: Úspěšnost koeficientů při použití Mahalanobisovy vzdálenosti (v %)

Metoda/koeficient	RMSSTD	CHF	PTS	D-B	Dunn
Nejbližšího souseda	5,26	47,37	52,63	52,63	47,37
Nejvzdálenějšího souseda	21,05	26,32	31,58	36,84	36,84
Centroidní metoda	0,00	52,63	42,11	63,16	36,84
Průměrná vzdálenost	5,26	47,37	52,63	63,16	52,63
Wardova metoda	26,32	42,11	21,05	5,26	57,89

Zdroj: vlastní zpracování

Jak vyplývá z tabulky 2, při užití Mahalanobisovy míry vzdálenosti je úspěšnost vybraných koeficientů opět nižší. Nejlepších výsledků bylo dosaženo opět při použití Daviesova-Bouldinova indexu za současné aplikace s metodou průměrné vzdálenosti či centroidní metody (63,16 %).

V tabulce č. 3 jsou uvedeny rozdíly v úspěšnosti jednotlivých koeficientů při jejich aplikaci s různými metodami shlukování a oběma měrami vzdáleností. Z tabulky vyplývá, že lepších výsledků je ve většině případů dosaženo při aplikaci Mahalanobisovy míry vzdálenosti, zejména při užití centroidní metody a metody průměrné vzdálenosti. Rozdíl v tomto případě činí 21,05 % ve prospěch Mahalanobisovy míry vzdálenosti.

Tabulka č. 3: Rozdíly v úspěšnostech koeficientů (Mahalanobisova – Euklidova vzdálenost) v %

Metoda/koeficient	RMSSTD	CHF	PTS	D-B	Dunn
Nejbližšího souseda	-5,26	0,00	10,53	15,79	-10,53
Nejvzdálenějšího souseda	-5,26	10,53	10,53	-26,32	-5,26
Centroidní metoda	-31,58	15,79	15,79	21,05	-10,53
Průměrná vzdálenost	-21,05	15,79	26,32	21,05	10,53
Wardova metoda	5,26	5,26	-15,79	-10,53	15,79

Zdroj: vlastní zpracování

5. ZÁVĚR

Shluková analýza je vícerozměrná statistická metoda, jejímž cílem je klasifikace objektů do skupin. Ke stanovení optimálního počtu shluků existuje mnoho způsobů (koeficientů), které je možné kombinovat s různými metodami a různými měrami vzdáleností.

Cílem tohoto článku bylo ukázat příklad stanovení počtu shluků u vybraných koeficientů, které jsou aplikovány do oblíbených softwarových produktů. K vyhodnocení schopnosti uvedených koeficientů správně stanovit počet shluků bylo provedeno mnoho analýz, jako například také v [5], [6]. V tomto článku byla podrobně analyzována skutečnost překrývajících se shluků na skutečných datových souborech z databáze *The UCI Machine Learning Repository*. V analýzách, které jsou dále uvedeny v [11] a [13], je provedeno srovnání na mnoho generovaných souborech, aby bylo možné stanovit srovnatelné podmínky a určit schopnost použití těchto koeficientů v různých podmínkách.

Na základě uvedených analýz je možné konstatovat, že u dobře separovaných shluků nemá počet výsledných shluků ani počet proměnných vliv na úspěšnost jednotlivých koeficientů. Lepších výsledků je obecně dosaženo při použití Euklidovy vzdálenosti, viz [13]. Čím je však separace shluků nižší, tím jsou koeficienty pro stanovení počtu shluků méně úspěšné. Jak bylo uvedeno výše, nejvyšší úspěšnost při analýze vybraných reálných datových souborů měl Daviesův-Bouldinův index, u kterého bylo při použití Mahalanobisovy míry vzdálenosti dosaženo vyšší úspěšnosti o 21,05 %. Použitelnost koeficientů pro stanovení optimálního počtu shluků v případě, že se shluky značně překrývají, byla velmi nízká. V takovémto případě je tedy lepších výsledků dosaženo při použití Mahalanobisovy míry vzdálenosti.

Poděkování

Tento článek byl vytvořen za podpory prostředků dlouhodobé institucionální podpory číslo IP400040 Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze.

LITERATURA

[1] CANNON, R. L. – DAVE, J. V. – BEZDEK, J. C.: Efficient Implementation of the Fuzzy c-means Clustering Algorithms. IEEE Transactions On Pattern Analysis and Machine Inteligence, 1989, No. 7, p. 773-781.

[2] DAVIES, D. L. – BOULDIN, D. W.: A Cluster Separation Measure. IEEE Transactions On Pattern Analysis and Machine Inteligence, 1979, No. 4, p. 224-227.

[3] DUNN, J.: Well Separated Clusters and Optimal Fuzzy Partitions. In: Journal of Cybernetics, 1974, No. 4, p. 95-104.
[4] GAN, G. – MA, CH. – WU, J.: Data Clustering Theory, Algorithms, and Applications. Philadelphia: ASA, 2007. ISBN: 978-0-898716-23-8.
[5] HALKIDI, M. – Vazirgiannis, M.: Clustering Validity Assessment: Finding the optimal partitioning of a data set. The Proceedings of ICDM. California, 2001, p. 1-9.
[6] HALKIDI, M. – BATISTAKIS, Y. – VAZIRGIANNIS, M.: On Clustering Validation Techniques. Journal of Intelligent Information System, 2001, No. 2-3, p. 107-145.
[7] HALKIDI, M. – Vazirgiannis, M. – BATISTAKIS, I.: Quality scheme assessment in the clustering proces. Proceedings of PKDD, 2000, p. 265-276.
[8] KOVÁCS, F. – LEGÁNY, C. – BABOS, A.: Cluster Validity Measurement Techniques. World Scientific and Engineering Academy and Society (WSEAS), 2006, p. 388-393.
[9] LÖSTER, T. – ŘEZANKOVÁ, H.: Evaluation of Clustering with Categorical and Mixed Type Variables and Cluster Number Determination. ISI 2011. Dublin, p. 1-6.
[10] LÖSTER, T.: Modification of CHF and BIC Coefficients for Evaluation of Clustering with Mixed Type Variables. In: Research Journal of Economics, 2013, No. 2, p. 1-4.
[11] LÖSTER, T.: The Evaluation of CHF coefficient in determining the number of clusters using Euclidean distance measure. The 8th International Days of Statistics and Economics. Praha, 2014, p. 858-869. ISBN 978-80-87990-02-5.
[12] LÖSTER, T.: Metody shlukové analýzy a jejich hodnocení. 1. vyd. Slaný: Melandrium, 2014. 132 s. ISBN 978-80-86175-88-1.
[13] LÖSTER, T.: The Evaluation of CHF coefficient in determining the number of clusters using Mahalanobis distance measure. 14th Conference on Applied Mathematics – Aplimat 2015 Bratislava, 2015, p. 546-554. ISBN 978-80-227-4314-3.
[14] MAULIK, U. – BANDYOPADHYAY, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2002, No. 12, p. 1650-1654.
[15] ŘEZANKOVÁ, H.: Hodnocení kvality shluků, Analýza dat 2008/II, TriloByte Statistical Software, Pardubice, 2009, s. 19–40. ISBN 978-80-904053-1-8.
[16] ŘEZANKOVÁ, H. – HÚSEK, D. – SNÁŠEL, V.: Shluková analýza dat. 2. rozšíř. vyd. Praha: PROFESSIONAL PUBLISHING, 2009. 218 s. ISBN 978-80-86946-81-8.
[17] ŘEZANKOVÁ, H. – ŽELINSKÝ, T.: Faktory míry materiální deprivace v České republice a jejich vztahy k typu domácnosti. In: Ekonomický časopis, 2014, č. 4, s. 394–410. ISSN 0013-3035.
[18] http://archive.ics.uci.edu/ml/datasets.html

RESUME

In case of the well-separated clusters, neither the number of the resulting clusters, nor the number of variables affect the efficiency of individual clusters. Better results are generally achieved by using the Euclidean distance. However, the lower the separation of clusters, the less efficient are the coefficients determining the number of clusters. The applicability of coefficients for determining the optimal number of clusters is very low if the clusters are significantly overlapped. In this case, better results are achieved by using Mahalanobis distance.

PROFESNÍ ŽIVOTOPIS

Ing. Tomáš Löster, Ph.D., působí na Fakultě informatiky a statistiky Vysoké školy ekonomické v Praze. Ve vědecko-výzkumné práci se zaměřuje na vícerozměrné statistické metody a statistické výpočetní prostředí. Největší pozornost ze statistických vícerozměrných metod je soustředěna na shlukovou analýzu. V této oblasti uchazeč publikoval řadu článků. Vyučuje předměty z oblasti statistiky a statistických metod. Je autorem či spoluautorem několika monografií, učebnic, skript a mnoha vědeckých článků. Působí v České statistické společnosti jako hospodář.

KONTAKT

tomas.loster@vse.cz