KLASIFIKACE V PROGRAMOVÝCH SYSTÉMECH PRO ANALÝZU DAT

HANA ŘEZANKOVÁ A DUŠAN HÚSEK

Abstrakt. The current classification methods are categorized. Their theoretical capabilities are discussed. The work includes the classification of cases, variables and categories. Moreover, implementation of these methods in six statistical packages (MINITAB, S-PLUS, SPSS, STATGRAPHICS, STATISTICA and SYSTAT) and their additional software systems (SPSS Answer Tree, STATISTICA Neural Networks) is described.

Проводится разбивка на категории существующих методов статистической классификации. Обсуждаются их теоретические основы. В работе рассматривается классификация случаев, переменных и категорий. Более того, дается описание использования етих методов в шести статистических пакетах прикладных программ (MINITAB, S-PLUS, SPSS, STATGRAPHICS, STATISTICA и SYSTAT) и в их дополнительных программных системах (SPSS Answer Tree, STATISTICA Neural Networks).

1. Úvod

Pojem *klasifikace* používaný při analýze dat je spojen s širokým okruhem metod, kterým věnují značnou pozornost jak statistikové, tak odborníci specializující se na takové oblasti jako jsou data mining, získávání znalostí z databází, včetně specialistů na neuronové sítě. V některých zemích existují odborné společnosti zabývající se touto tématikou, které jsou sdruženy do mezinárodní společnosti *International Federation of Classification Societies* (IFCS). Pod záštitou IFCS se konají konference *Data Science, Classification, and Related Methods*. Existuje také specializovaný časopis *Journal of Classification*.

Klasifikační metody můžeme charakterizovat následujícím způsobem. Sledujeme určité objekty, které se více či méně navzájem odlišují, takže může existovat několik skupin těchto objektů. Cílem je zařadit buď některé z objektů nebo všechny objekty do skupin.

2. Klasifikace klasifikačních metod

Metody obsažené v programových systémech můžeme roztřídit podle různých hledisek. Dále budou uvedena některá z nich, budou specifikovány příslušné skupiny metod, případně bude uveden konkrétní postup, který je buď jediným nebo charakteristickým zástupcem dané skupiny. U členění podle třetího hlediska pak budou vyjmenovány jednotlivé metody a postupy, které můžeme nalézt v programových systémech pro analýzu dat.

2000 *Mathematics Subject Classification.* Primary 62H30.

Klíčová slova. Klasifikační metody.

Tato práce vznikla v rámci grantu No.LN00B096 – Centrum aplikované kybernetiky.

1. Jedním z hledisek je *předmět klasifikace*. Při analýze dat můžeme rozlišit následující předměty klasifikace:
   1.1. Objekty, tj. statistické jednotky
   1.2. Proměnné, tj. statistické znaky
   1.3. Kategorie proměnných
      1.3.1. Jedné proměnné (např. shluková analýza)
      1.3.2. Dvou proměnných (dvourozměrná shluková analýza, korespondenční analýza)
      1.3.3. Více proměnných (optimální škálování)
2. Jiné hledisko může sledovat, *kdy a jakým způsobem je stanoven počet skupin*. Podle něho můžeme rozlišit následující kategorie:
   2.1. Počet skupin musí být stanoven před analýzou, jejímž cílem je klasifikace
      2.1.1. Dáno názorem uživatele, který analyzuje data (nehierarchická shluková analýza)
      2.1.2. Dáno počtem hodnot vysvětlované proměnné (diskriminační analýza). V této skupině můžeme dále rozlišit, zda vysvětlovaná proměnná je dichotomická (logistická regresní analýza), nominální (multinomická logistická regrese) či ordinální (ordinální regrese), u které se bere v úvahu, zda jsou kategorie rovnoměrně zastoupeny, nebo jsou více zastoupeny nižší či vyšší hodnoty, případně jsou více zastoupeny extrémní hodnoty.
   2.2. Počet skupin je zjišťován analýzou, jejíž cílem je klasifikace
      2.2.1. Metodou může být navržen konkrétní počet (faktorová analýza)
      2.2.2. Počet stanovuje uživatel na základě výsledků analýzy (hierarchická shluková analýza)
3. Nejčastěji používaným hlediskem je zřejmě to, *zda se můžeme při klasifikaci řídit nějakým vzorem* (existuje „učitel“) *nebo ne*. Podle toho rozlišujeme
   3.1. Učení s učitelem (supervised learning), které se vztahuje pouze na klasifikaci objektů a známý počet skupin (na základě vzoru je vytvořen model, který umožňuje zařazování objektů do skupin); slouží k odhadu hodnoty vysvětlované proměnné, která je kategoriální. V literatuře zaměřující se na data mining jsou jako klasifikační označovány pouze tyto metody, jde tedy o *klasifikaci v užším významu*:
      3.1.1. Diskriminační analýza
      3.1.2. Zobecněný lineární model - GLM (Generalized Linear Model)
         3.1.2.1. Logistická regresní analýza
         3.1.2.2. Multinomická logistická regresní analýza
         3.1.2.3. Ordinální regresní analýza
      3.1.3. Kategoriální regresní analýza
      3.1.4. Klasifikační stromy
         3.1.4.1. Metoda CHAID (Chi-squared Automatic Interaction Detection) - vysvětlovaná proměnná může být jak nominální, tak ordinální (metoda je používána i v případě spojité vysvětlované proměnné)
         3.1.4.2. Metoda Exhaustive CHAID

3.1.4.3. Metoda C&RT (Classification and Regression Trees) - vysvětlovaná proměnná může být jak nominální, tak ordinální (metoda je používána i v případě spojité vysvětlované proměnné)
3.1.4.4. Metoda QUEST (Quick, Unbiased, Efficient Statistical Tree) - lze použít pouze pro nominální vysvětlovanou proměnnou
3.1.4.5. Další metody (CART, CLS, ID3, C4.5, AID, TREEDISC)
3.1.5. Neuronové sítě
3.1.5.1. MLP (MultiLayer Perceptrons) - vícevrstvý perceptron
3.1.5.2. RBF (Radial Basis Functions) - radiální bazické funkce
3.1.5.3. PNN (Probabilistic Neural Networks) - pravděpodobnostní neuronové sítě
3.1.5.4. LNN (Linear Neural Networks) - lineární neuronové sítě
3.1.5.5. LVQ (Learning Vector Quantization) - vektorová kvantizace
3.1.6. GUHA[1]

3.2. Učení bez učitele (unsupervised learning), které zahrnuje jednak shlukování či segmentaci (objektů, proměnných i kategorií), jednak redukci dat (proměnných či kategorií). V literatuře zaměřující se na data mining se tyto metody neoznačují jako klasifikační, ale spadají do skupiny postupů, jejichž cílem je *shlukování*, případně segmentace:
3.2.1. Shluková analýza
3.2.1.1. Hierarchická shluková analýza
3.2.1.2. Nehierarchická shluková analýza
3.2.1.3. Dvourozměrná shluková analýza (two-way joining)
3.2.2. Faktorová analýza
3.2.3. Vícerozměrné škálování
3.2.4. Korespondenční analýza
3.2.5. Optimální škálování (analýza homogenity, kategoriální analýza hlavních komponent)
3.2.6. Neuronové sítě
3.2.6.1. AR (Adaptive Resonance Theory)
3.2.6.2. KFM (Kohonen Feature Maps) - Kohonenovy mapy

[1]K metodám klasifikace můžeme dále zařadit metody, které *analyzují vztahy mezi kategoriemi* různých proměnných. Základem je asociační algoritmus pro odvozování pravidel typu If X, then Y, tj. implikace typu IF (logická kombinace fakt) THEN fakt, přičemž *fakt* je elementární logický výrok. Je zjišťováno, kolik procent z určité logické kombinace fakt (antecedentu) implikuje fakt na pravé straně pravidla (succedent) a kolik procent záznamů se vyskytuje v této asociaci. Příkladem programového systému, který je určen pro výše uvedené analýzy, je např. GUHA +- (General Unary Hypotheses Automaton). Na základě analyzovaných vztahů lze pro určitou kombinaci kategorií některých vysvětlujících proměnných předpovědět hodnotu (kategorii) vysvětlované proměnné (obdobně jako při použití klasifikačních stromů). Metodu lze tedy využít pouze pro kategorizované proměnné.

3.2.6.3. HNN (Hopfield like Neural Network) - síť Hopfieldova typu

3.2.7. Genetické algoritmy

3. Porovnání možností statistických programových systémů v oblasti klasifikace

Sledovány byly vybrané možnosti šesti statistických programových systémů, kterými jsou MINITAB 13 (demo verze), S-PLUS 4.5, SPSS 10.0, STATGRAPHICS *Plus* 4.0 (demo verze), STATISTICA 5.1 a SYSTAT 9.0 (demo verze), a jich rozšiřujících systémů (SPSS Answer Tree 2.0, STATISTICA Neural Networks 3.0). Metody klasifikace byly rozděleny do dvou základních skupin, které lze charakterizovat jako
- předpovídání hodnot kategoriální vysvětlované proměnné a
- shlukování.

Do podrobnějšího hodnocení v rámci první skupiny nebyl zahrnut systém S-PLUS, neboť neobsahuje diskriminační analýzu a při logistické regresní analýze nabídkový režim neposkytuje většinu ze sledovaných možností. Zjištěné skutečnosti jsou uvedeny v tabulkách 1 a 2, sledovány jsou následující možnosti.

Možnost	Vysvětlení
Výběr prom.	nabídka metod, které umožňují vybrat z množiny vysvětlujících proměnných takovou podmnožinu, jež nejlépe vysvětluje hodnoty vysvětlované proměnné (obvykle jsou nabízeny postupy forward a backward)
Váhy dle počtu	jednotlivým skupinám lze přiřadit váhy dle počtu objektů v dané skupině
Váhy dle uživat.	jednotlivým skupinám lze přiřadit váhy dle uživatele
Tab. úspěšnosti	výstup obsahuje tabulku sdružených četností pro skutečné a předpovězené hodnoty vysvětlované proměnné
- cross validation	známý objekt, pro který chceme získat předpověď, není zahrnut do analýzy
Předpovědi	způsob záznamu předpovědí hodnot vysvětlované proměnné
- text. výstup	předpovědi jsou zobrazeny jako textový výstup
- výstupní tab.	předpovědi jsou zobrazeny ve výstupní tabulce
- do dat. editoru	předpovědi jsou hodnoty nové proměnné, která je přidána do datového editoru ke zdrojovým datům
- nový soub.	předpovědi jsou uloženy do nového datového souboru
- pro zadané h.	předpovědi jsou počítány pro nový případ (zadaný vektor hodnot)
- pravděp.	jako výsledky jsou uváděny pravděpodobnosti
Grafy	počet nabízených typů grafů
Překód. Y (0 a 1)	vysvětlovaná proměnná může nabývat jiných hodnot než 0 a 1
- automat.	překódování provádí systém v případě potřeby automaticky
Kód. kat. prom.	kategoriální proměnná nabývající k kategorií je převedena na k-1 pomocných proměnných
Mezní hodnota	lze zadat hodnotu od 0,01 do 0,99 - jestliže je vypočítána hodnota větší než zadaná mezní, je předpověď vysvětlované proměnné rovna jedné
Úspěšnost	způsob posuzování úspěšnosti při předpovědích

Co se týče grafů, které ilustrují výše uvedené analýzy, pak stejný počet grafů neznamená stejné typy. Obvykle se grafické výstupy u jednotlivých programových systémů poněkud liší.

V systému *S-PLUS*, který nebyl do výše uvedeného přehledu zahrnut, je k dispozici logistická regresní analýza (předpovědi lze uložit do nového datového souboru, nabídka obsahuje 8 typů grafů) a Poissonova regresní analýza. Součástí systému jsou dále klasifikační a regresní stromy, které zahrnují 2 metody (jednu pro kvalitativní, druhou pro kvantitativní vysvětlovanou proměnnou). Neuronové sítě obsaženy nejsou.

Pokud jde o programové produkty umožňující provádět analýzu dat pomocí neuronových sítí, pak autoři měli k dispozici pouze *STATISTICA Neural Networks*. Tento systém poskytuje pro předpovídání hodnot kategoriální vysvětlované proměnné 4 typy neuronových sítí, které jsou ve druhé kapitole označené jako 3.1.5.1 až 3.1.5.4. Pro MLP (vícevrstvý perceptron) je k dispozici 5 algoritmů: back propagation (zpětné šíření), conjugate gradient descent algoritmus, Levenbergův-Marquardtův algoritmus, quick propagation (rychlé šíření) a Delta-bar-Delta propagation. Možnosti jsou tedy poměrně značné. Binární proměnné musí být kódovány číslicemi 0 a 1, produkt poskytuje tabulku úspěšnosti a předpovědi. Zvláštností systému je, že tabulka úspěšnosti zahrnuje kromě statistických jednotek zařazených do daných kategorií také statistické jednotky nezařazené.

Tab. 1 Předpovídání hodnot kategoriální vysvětlované proměnné

	MINITAB	SPSS	STAT GRAPHICS	STATISTICA	SYSTAT
Diskriminační analýza					
Výběr prom.	Ne	5 metod	Forward, backw.	Forward, backw.	Forward, backw.
Váhy dle počtu	Ne	Ano	Ano	Ano	Ne
Váhy dle uživat.	Ano	Ne	Ano	Ano	Ne
Tab. úspěšnosti	Ano	Ano	Ano	Ano	Ano
- cross validation	Ano	Ano	Ne	Ne	Ano
Předpovědi	Do dat. editoru Text. výst. (pro zadané hodnoty)	Do dat. editoru	Text. výst.	Výst. tab. Nový soub.	Nový soub.
Grafy	0	3	3	3	0
Logistická regresní analýza (binární)					
Překód. Y (0 a 1)	Automat.	Automat.	Ne	Lze zadat	Automat.
Výběr prom.	Ne	3 forw., 3 backw.	Forward, backw.	Ne	Ne
Kód. kat. prom.	Ano	Ano	Ano	Ne	Ano
Mezní hodnota	Ne	Ano	Ano	Ne	Ne
Úspěšnost	Chí-kvadrát test	Tabulka úspěšnosti	Chí-kvadrát test	Ne	Tabulka úspěšnosti
Předpovědi	Ne	Do dat. editoru	Ne	Ne	Nový soub. (pravděp.)
Grafy	8	1	7	7	0
Klasifikační stromy					
Produkt	Ne	Answer Tree	Ne	Součást paketu	Součást paketu
Možnosti	x	4 metody	x	3 metody	6 funkcí
Neuronové sítě					
Produkt	Ne	Neural Connection	Ne	Neural Networks	Ne

Poznámky. Produkt *Answer Tree* zahrnuje metody označené v kapitole 2 jako 3.1.4.1 až 3.1.4.4. Kvalita modelu je posuzována na základě tabulky, která je obdobou tabulky úspěšnosti, avšak celkově je charakterizována podílem chybných předpovědí. V systému *SYSTAT* lze v rámci klasifikačních stromů vybírat z šesti typů ztrátových funkcí.
MINITAB umožňuje kromě výše uvedených analýz provádět též nominální a ordinální regresní analýzu, avšak neposkytuje žádné ze sledovaných možností.
Systém*SPSS* nabízí kromě dvou výše uvedených postupů také kategoriální regresní analýzu. Některé možnosti jsou porovnány v tabulce 2.

Tab. 2 Další možnosti systému SPSS z oblasti regresní analýzy (RA)

	Multinomická RA (nominální)	*Ordinální RA*	*Kategoriální RA*
Výběr prom.	Ne	Ne	Ne
Tab. úspěšnosti	Ano	Ne	Ne
Předpovědi	Ne	Ano	Ne
Graf	Ne	Ne	Ano

Pokud jde o oblast *shlukování*, porovnání je uvedeno v tabulce 3. Je založeno především na možnostech nabídkových režimů, ve dvou případech jsou též zmíněny možnosti dostupné pomocí příkazů, pomocí nichž lze dále získat například více typů grafů (SPSS). Byly sledovány následující možnosti:

Možnost	Vysvětlení
Faktorová analýza	
Metody extrakce	počet metod extrakce
Metody rotace	počet metod rotace
Grafy	počet nabízených typů grafů
Hierarchická shluková analýza	
Z matice vzdál.	vstupem pro analýzu může být matice vzdáleností
Shluk. prom.	lze provádět shlukování proměnných (sloupců v datové matici)
Standard. hodnot	lze zadat transformaci vstupních hodnot – standardizaci (obvykle odečtení aritmetického průměru a vydělení směrodatnou odchylkou, v systému S-PLUS je místo směrodatné odchylky používána absolutní odchylka), příp. jiné transformace (např. převedení do intervalu od = -1 do 1 či od 0 do 1)
Míry vzdál.	počet měr vzdálenosti
Míry podob.	počet měr podobnosti
Míry nepod.	počet měr nepodobnosti
Míry pro bin. pr.	počet měr pro binární proměnné
Transf. měr	lze zadat transformaci měr (absolutní hodnoty, převedení do int. od 0 do 1)
Aglomer. metody	počet aglomerativních postupů
Metody dělení	počet metod pro dělení shluků
Přísl. ke shlukům	způsob záznamu příslušnosti ke shlukům (možnosti viz Předpovědi v předcházející skupině metod), v SPSS lze příslušnosti zaznamenat do datového editoru pouze v případě, že jsou shlukovány objekty
- tab. v text. výst.	příslušnosti jsou zaznamenány do tabulky, která je součástí textového výstupu
Icicle graf (banner plot)	směr kreslení grafu (pokud je obsažen) - vertikální nebo horizontální
Dendrogram (tree plot)	směr kreslení dendrogramu - vertikální nebo horozontální
Nehierarchická shluková analýza	
Metody	konkrétní metoda (k-means = k-průměrů) nebo počet metod

Tab. 3 Shlukování

	MINITAB	S-PLUS	SPSS	STAT GRAPHICS	STATISTICA	SYSTAT
Faktorová analýza						
Metody extrakce	2	2	7	2	6	3
Metody rotace	4	12	5	3	8	5
Grafy	3	2	2	2	2	2
Hierarchická shluková analýza						
Z matice vzdál.	Ano	Ano	Příkaz	Ne	Ano	Ano
Shluk. prom.	Ano	Ne	Ano	Ano	Ano	Ano
Standard. hodnot	1	1(+ 3)	6	1	0	0
Míry vzdál.	5	2	5	3	6	8
Míry podob.	2	0	2	0	1	0
Míry nepod.	0	Příkaz	2	0	0	0
Míry pro bin. pr.	0	2	26	0	0	0
Transf. měr	0	0	3	0	0	0
Aglomer. metody	7	5	7	6	7	6
Metody dělení	0	2	0	0	0	0
Přísl. ke shlukům	Do dat. editoru	Nový soub.	Tab.výst. v text. výst. Do dat. editoru	Text. Do dat. editoru	Ne	Text. výstup Do dat. editoru
Icicle graf (banner plot)	Ne	Horiz.	Vert. + horiz.	Ne	Ne	Ne
Dendrogram (tree plot)	Vert.	Vert.	Horiz.	Vert. + vert.	Horiz.	Ne
Jiné grafy	0	0	0	3	0	Polární
Nehierarchická shluková analýza						
Metody	*k*-means	3 (vč. fuzzy)	*k*-means	*k*-means (i prom.)	*k*-means (i prom.)	*k*-means (8 měr)
Jiné techniky shlukové analýzy						
	Ne	Ne	Ne	Ne	Dvourozm.	Additive tree
Vícerozměrné škálování						
	Ne	Ne	Ano	Ne	Ano	Ano
Korespondenční analýza						
	Ne	Ne	Ano	Ne	Ano	Ano
Analýza homogenity						
	Ne	Ne	Ano	Ne	Ano	Ne
Kategoriální analýza hlavních komponent						
	Ne	Ne	Ano	Ne	Ne	Ne

Poznámky.

V *S-PLUS* je při shlukové analýze jiná nabídka pro transformaci vstupních hodnot u konkrétních analýz (pouze standardizace) a jiná u výpočtu matice vzdáleností (3 další možnosti - podle typů proměnných).

V systémech *MINITAB* a *SYSTAT* jsou do měr vzdáleností zahrnuty i transformované míry podobnosti (Pearsonův korelační koeficient). *SYSTAT* poskytuje míry vzdálenosti pro různé typy dat - intervalová, ordinální a nominální data a četnosti.

Systémy *STATGRAPHICS* a *STATISTICA* umožňují aplikovat algoritmus *k*-průměrů také

na proměnné, *STATGRAPHICS* a *SYSTAT* poskytují v rámci tohoto shlukování zvolit míru vzdálenosti.

Pokud jde o systém *STATISTICA Neural Networks*, pak z typů sítí uvedených v 3.2.6 tento produkt poskytuje pouze Kohonenovy mapy. Pro shlukování a redukci dat však mohou být využity rovněž neuronové sítě uvedené v 3.1.5, například lineární neuronové sítě se používají pro analýzu hlavních komponent. Pomocí radiálních bazických funkcí lze provádět shlukovou analýzu odpovídající algoritmu k-průměrů, k dispozici je též algoritmus k-nejbližšího souseda. K výběru podmnožiny vhodných vysvětlujících proměnných lze použít genetické algoritmy.

4. Závěr

V příspěvku byla věnována pozornost vybraným možnostem některých programových systémů pro analýzu dat. U základních metod byl sledován především komfort nabízený uživateli z hlediska požadavků na vstupní data, například možnost překódování hodnot dichotomické vysvětlované proměnné či převedení nominální vysvětlující proměnné na pomocné proměnné přímo v rámci klasifikační procedury, a dále z hlediska poskytovaných výstupů, např. zobrazení tabulky úspěšnosti a předpovědí hodnot vysvětlované proměnné. Důležité jsou též grafy, které hrají významnou úlohu při snadné interpretaci výsledků.

Jestliže programový systém tyto možnosti neposkytuje, uživatel musí získávat dodatečné informace jinými způsoby. Například uživateli nejde obvykle o to, aby si vytiskl parametry modelu, ale o to, aby mohl zařadit neznámý objekt do některé ze známých skupin objektů, což některé procedury bohužel neumožňují.
V oblasti shlukování bylo sledováno spíše množství různých měr a algoritmů. Z hlediska měr podobností či nepodobností je třeba vyzdvihnout systém *SPSS* s jeho 26 mírami pro binární proměnné. Je však otázkou, kolik procent uživatelů některou z těchto měr využije a zda použije míru správnou, neboť ani manuál ani nápověda neposkytuje základní informace o typech binárních proměnných (symetrických a asymetrických).

Existuje samozřejmě celá řada dalších faktorů, podle kterých lze hodnotit programové systémy. Kromě snadnosti ovládání je to například možnost výpočtů s daty obsahující chybějící údaje. V odborných kruzích jsou preferovány systémy *S-PLUS* a *SPSS*, které poskytují jednak rozsáhlou nabídku metod, jednak možnost programování. Ve speciálních případech však můžeme zjistit, že ani jeden z nich nelze použít.

Jestliže například chceme shlukovat asymetrické binární proměnné, matice vzdáleností může obsahovat chybějící hodnoty i v případě, kdy se ve zdrojových datech chybějící údaje nevyskytují. *S-PLUS* hlásí, že se nesmí vyskytovat žádná chybějící hodnota, pro *SPSS* je na závadu, když chybějících hodnot je mnoho. Řešením je v tomto případě vypočítat matici vzdáleností pomocí některého z těchto produktů (*SPSS* je přijatelnější z hlediska ovládání) a shlukování nechat provést v systému *STATISTICA*. Postup je sice poněkud komplikován tím, že systém *STATISTICA* vyžaduje matici vzdáleností ve speciálním tvaru (musí obsahovat navíc 3 řádky, které charakterizují soubor), nicméně je zřejmě jediným řešením dané úlohy.

Z hlediska množství typů metod využitelných pro úlohy klasifikace lze nejlépe hodnotit systém *SPSS*, za nímž následují systémy *STATISTICA* a *SYSTAT*. Nejméně typů metod zahrnuje *STATGRAPHICS*. Jednoznačně však nelze doporučit žádný systém, neboť každý ze sledovaných zahrnuje buď určitý postup či speciální grafy, které v jiných systémech zahrnuty nejsou. Jako příklady lze uvést speciální postupy

regresní analýzy a speciální míry vzdáleností v *SPSS*, dvourozměrnou shlukovou analýzu v systému *STATISTICA*, metodu additive tree joining a polární graf u shlukové analýzy v systému *SYSTAT*, Poissonovu regresní analýzu a speciální postupy ve shlukové analýze (fuzzy přiřazování do shluků a postupy pro dělení shluků) v *S-PLUS*.

V současné době již výše uvedené programové systémy poskytují nabídkový režim pro snadné ovládání. Bohužel u některých produktů (*S-PLUS, SPSS*) nezahrnuje tento nabídkový režim všechny analytické možnosti systému a dokonce některé základní možnosti musí být někdy zadávány pomocí příkazů (2D graf u faktorové analýzy v *SPSS*).

Pro každý typ analýzy je nezbytnou součástí práce s daty, přičemž základní možností je výběr proměnných pro analýzu. Z tohoto hlediska je dle našeho názoru nejméně přizpůsoben potřebám běžného uživatele systém *S-PLUS*. Obtížnější na ovládání jsou pro uživatele statistických programových systémů též speciální produkty jako *Neural Networks*, neboť vyžadují specifický způsob zadávání.

Jak potvrzují diskuse s kolegy, kteří se zabývají analýzou dat, obvykle nelze vystačit s jediným programovým systémem, je nutné mít nejméně dva a při výpočtech je kombinovat. Pokud lze výsledky získat prostřednictvím různých produktů, měly by být porovnány, neboť softwarové produkty se mohou lišit z hlediska spolehlivosti výpočtů.

Literatura

[1] *Answer Tree 2.0 User's Guide*. SPSS Inc., Chicago 1998.

[2] Bigus, J. P.: *Data Mining with Neural Networks*. McGraw-Hill, 1996.

[3] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi. A.: *Discovering Data Mining: From Concept to Implementation*. Prentice Hall PTR, New Jersey 1998.

[4] Groth, R.: *Data Mining: A Hands-On Approach for Business Professionals*. Prentice Hall PTR, New Jersey 1998.

[5] Hartigan, J. A.: *Clustering Algorithms*. John Wiley & Sons, New York 1975.

[6] Húsek, D., Řezanková, H., Frolov A. A.: Neural Networks Capable of Boolean Factor Analysis. In: *COMPSTAT 2000. Proceedings in Computational Statistics - Short Communications and Posters* (Ed. Jansen, W., Bethlehem, J. G.). Statistics Netherlands, Voorburg 2000, 41-42.

[7] Klaschka, J., Antoch, J.: Jak rychle pěstovat stromy. *ROBUST'96*. JČMF, Praha 1997, 91-106.

[8] Řezanka, T., Řezanková, H.: Characterization of fatty acids and triacylglycerols in vegetable oils by gas chromatography and statistical analysis. *Analytica Chimica Acta*, **398** (1999), 253-261.

[9] Řezanková, H.: Cluster Analysis Algorithms in Software Systems. In: *Socio-Economical Applications of Statistical Methods* (Ed. Ostasiewicz, W.). University of Economics, Wroclaw 2000, 154-162.

[10] Řezanková, H., Húsek, D.: Metody pro redukci znaků sledovaných při analýze dat. *8. mezinárodný seminár Výpočtová štatistika*, SŠDS, Bratislava 1999, 84-87.

[11] Řezanková, H., Húsek, D.: Modeling technique for Data Mining. *Acta Oeconomica Pragensia*, **8** (2000), No. 3, 125-132.

[12] *SPSS Advanced Models 10.0*. SPSS Inc., Chicago 1999.

[13] *SPSS Categories*. SPSS Inc.

[14] *STATISTICA*. StatSoft Inc, Tulsa1998

[15] *STATISTICA Neural Networks*. StatSoft Inc., Tulsa 1998

[16] *S-PLUS 4 Guide to Statistics*. MathSoft Inc., Seattle 1997.

[17] Tvrdík, J.: Logistická regrese a vyhledávání modelů. *ROBUST'98*, JČMF, Praha 1998, 187-194.

VŠE, KSTP, nám. W. Churchila 4, 130 00 Praha 3
E-mail: rezanka@vse.cz

ÚI, AV ČR, Pod Vodárenskou Věží 2, 182 07 Praha
E-mail: dusan@cs.cas.cz