SHLUKOVACÍ METODY V DATA MININGU

E + M

Petr Klímek

1. Data Mining

Obecná definice data mining popisuje jako proces výběru, prohledávání a modelování ve velkých objemech dat, sloužící k odhalení dříve neznámých vztahů mezi daty, za účelem získání konkurenční výhody. Problematikou data miningu se zabývá práce [1], dále například v časopisu E+M Ekonomie a Management [2].

DM bývá v širším slova smyslu, např. v [5], definován jako získávání dosud neznámých, ověřených a použitelných znalostí z rozsáhlých databází pro provádění klíčových manažerských rozhodnutí. Příklady obchodního zadání, které mohou vést k zavádění technik DM, jsou:

- Klasifikace (např. „Představuje toto hlášení o škodní události pojišťovací podvod?“);
- Odhad (např. „Jaká je obchodní hodnota zákazníka?“);
- Předpovídání (např. „Kteří zákazníci od nás pravděpodobně odejdou v průběhu nejbližších šesti měsíců?“);
- Asociační analýza (např. „Které produkty se obvykle kupují společně?“);
- Seskupování podle podobností (např. „Které skupiny zákazníků mají nějaké společné charakteristiky?“);
- Deskripce (např. „Které atributy nejvíce charakterizují chování určité skupiny zákazníků?“).

V praxi může nastat mnoho dalších situací či problémů, k jejichž řešení jsou vhodné metody data miningu, např. v [7] nebo v [8]. Využití data miningových metod v ČR podrobně rozebírá [3]. V tomto článku se dále budeme zabývat předposlední skupinou technik, a to seskupováním podle podobností neboli shlukováním.

2. Shluková analýza

Ve statistické literatuře [6] a praktických aplikacích známe celou řadu přístupů k řešení problémů dělení na shluky, z nichž jenom část je založena na statistických principech a používá statistickou interpretaci. Tyto přístupy vycházejí z kritéria, že všechny prvky se přiřazují do shluků tak, aby rozptyl uvnitř shluků byl minimální a současně rozptyl mezi shluky byl maximální.

Shluková analýza je souhrnný název pro řadu výpočetních postupů, jejichž cílem je rozklad daného souboru na několik relativně homogenních podsouborů (shluků) a to tak, aby jednotky (objekty) uvnitř jednotlivých shluků si byly co nejvíce podobné a jednotky (objekty) patřící do různých shluků si byly podobné co nejméně. Přitom každá jednotka je popsána skupinou znaků (proměnných).

Výsledek analýzy závisí na volbě proměnných, zvolené míře vzdálenosti mezi objekty a shluky a na zvoleném algoritmu výpočtu. Obecně lze úlohu shlukové analýzy vyjádřit takto: Máme n jednotek (objektů) a každá jednotka je charakterizována p znaky. Výsledky pozorování tvoří matici $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, kde $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ jsou p - členné vektory pozorování. Úkolem shlukové analýzy je rozdělení množiny $\mathbf{X}$ do množiny S; $S = \{S_1, S_2, ..., S_m\}$, kde $S_1, S_2, ..., S_m$ je m shluků, do nichž je provedeno seskupení objektů $\mathbf{x}_i$. Obecně může počet shluků dosahovat čísla n, praktický význam má však pouze takový počet shluků, který je podstatně menší než počet původních jednotek. Při tom se zpravidla požaduje, aby jednotlivé shluky byly navzájem disjunktní. Vytvořené shluky by měly být pokud možno kompaktní a navzájem relativně izolované, což je třeba vhodným způsobem kvantifikovat.

Úlohou shlukové analýzy je tedy na základě hodnot pozorovaných proměnných rozdělit soubor objektů do relativně homogenních skupin; objekty ve skupině jsou z hlediska těchto proměnných podobné a od objektů v jiných skupinách se liší. Shluková analýza se tak uplatňuje především při klasifikaci objektů, ale lze ji použít také k redukci počtu proměnných, a tedy ke zjednodušení úlohy (několik uvažovaných proměnných je nahrazeno jednou). Existence skupin podobných objektů je možné využít při organizování co nejvýhodnějšího postupu výběrového zjišťování. K hodnocení podobnosti (resp. nepodobnosti) objektů se nej-

častěji používají míry vzdáleností, a to například eukleidovská, Hammingova, Čebyševova nebo Mahalanobisova. Vzdálenosti dvojic jednotek lze uspořádat do čtvercové symetrické matice s nulami na hlavní diagonále. Při vlastním shlukování pak mohou být pro určení vzdálenosti dvou shluků použita různá kritéria: minimální vzdálenost dvojice objektů, jejich maximální vzdálenost, průměrná vzdálenost apod. Rozhodnutí o konečném počtu shluků vychází jak z teoretických (vzdálenosti shluků), tak i praktických hledisek (například cílový počet skupin je součástí klientovy objednávky).

Shlukovou analýzu lze velmi dobře uplatnit při segmentaci trhu, při sdružování zákazníků podle rozdílného kupního chování, je možné vytvářet skupiny srovnatelných obcí s cílem uplatnění obdobné marketingové strategie atd.

3. Míry vzdálenosti mezi objekty

Nejčastěji se k určení nepodobnosti dvou objektů používají následující míry nepodobnosti - vzdálenost.

A. Euklidovská vzdálenost objektů X_i a X_j je definována vztahem

$$d_1(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \qquad (1)$$

kde

x_{ik} je hodnota k-té proměnné u i-tého objektu,

x_{jk} je hodnota k-té proměnné u j-tého objektu.

Výhodou této míry je její výpočetní jednoduchost. Má však i některé nedostatky. Předpokládá totiž nekorelovanost proměnných, což je předpoklad, který je v praktických podmínkách obtížně splnitelný. Dále je značně závislá na měřítku proměnných, takže je vhodné pracovat s proměnnými v normovaném tvaru, tj. takovými, které mají nulový průměr a jednotkový rozptyl - jsou tedy bezrozměrnými čísly.

B. Hammingova vzdálenost objektů X_i a X_j je definována vztahem

$$d_2(X_i, X_j) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|. \qquad (2)$$

C. Mahalanobisova vzdálenost objektů X_i a Xj je definována vztahem

$$d_3(X_i, X_j) = (\mathbf{x_i} - \mathbf{x_j})^T \mathbf{C}^{-1} (\mathbf{x_i} - \mathbf{x_j}), \qquad (3)$$

kde $\mathbf{x_i}$ a $\mathbf{x_j}$ jsou p-členné vektory proměnných u i-tého a j-tého objektu, **C** je kovarianční matice.

Tato míra vzdálenosti přihlíží k vzájemným interkorelacím proměnných. Pokud jsou proměnné nekorelovány (párové korelační koeficienty jsou nulové), je Mahalanobisova vzdálenost rovna čtverci euklidovské vzdálenosti.

4. Shlukovací metody

V první fázi shlukování vypočteme vzájemné vzdálenosti všech objektů a zapíšeme je do matice. Tím dostaneme čtvercovou symetrickou matici $\mathbf{D} = \{d_{ij}\}$, která má na hlavní diagonále nuly. Na základě matice vzdáleností může být zahájena druhá fáze výpočtů, vlastní shlukovací procedura. Byla zkonstruována celá řada metod na určování vzdáleností mezi shluky; mezi nejznámější patří metody nejbližšího a nejvzdálenějšího souseda, metoda průměrné vzdálenosti a centroidní metoda.

Pro všechny tyto metody je společné, že na počátku tvoří každý objekt svůj vlastní shluk. V následujících krocích spojujeme vždy dva shluky, jejichž vzdálenost je minimální. Míra vzdálenosti mezi shluky závisí na zvolené metodě shlukování. Všechny shlukovací postupy končí po (n - 1) krocích tím, že všechny objekty splynou v jediný shluk. Při shlukování se za shluky považují i jednotlivé objekty.

Nechť $d_i(S_h, S_k)$ značí míru vzdálenosti mezi dvěma shluky S_h a S_k; vyslovíme následující definice.

A. Metoda nejbližšího souseda

Metoda nejbližšího souseda definuje vzdálenost dvou shluků jako vzdálenost jejich nejbližších prvků, tedy

$$d_1(S_h, S_k) = \min_{X_i \in S_h \in S_k} \{d(X_i, X_j)\}. \qquad (4)$$

B. Metoda nejvzdálenějšího souseda

Metoda nejvzdálenějšího souseda definuje vzdálenost dvou shluků jako vzdálenost jejich nejodlehlejších prvků, tedy

$$d_2(S_h, S_k) = \max_{X_i \in S_h \in S_k} \{d(X_i, X_j)\} \qquad (5)$$

C. Metoda průměrné vzdálenosti

Metoda průměrné vzdálenosti definuje vzdálenost dvou shluků jako průměrnou vzdálenost všech dvojic objektů z obou shluků, tedy

D. Metoda centroidní

$$d_3(S_h, S_k) = \frac{1}{n_h n_k} \sum_{X_i \in S_h} \sum_{X_j \in S_k} d(X_i, X_j). \quad (6)$$

Metoda centroidní definuje míru vzdálenosti mezi shluky jako vzdálenost jejich průměrů, tedy

$$d_4(S_h, S_k) = d(\bar{x}_h, \bar{x}_k), \quad (7a)$$

kde $\bar{x}_h = \frac{1}{n_h} \sum_{X_i \in S_h} X_i$, je počet objektů ve shluku. (7b)

Podrobnější výklad by přesáhl rámec problematiky. Statistické pakety, např. Statgraphics 5.0 a další obsahují potřebné programy pro provádění shlukové analýzy. Při shlukové analýze lze také použít metod tzv. fuzzy shlukování. Touto metodou se zabývá např. [4]. Dále se také nabízí aplikace metod, které vycházejí z biologicky inspirovaných algoritmů, které jsou popsány v poslední kapitole. [6]

5. Praktické příklady použití shlukové analýzy

Příklad 1 - jednorozměrná shluková analýza

V následujícím výzkumu máme za cíl rozlišit skupiny zákazníků s různým postojem k nakupování. Bylo vyzváno 20 respondentů, aby vyjádřili na sedmibodové škále stupeň svého souhlasu s následujícími výroky (1 = vůbec nesouhlasím, 7 = zcela souhlasím):

X1: O nákupy mám zájem. X2: Nákupy stojí příliš mnoho peněz. X3: Při návštěvě města vždy nakupuji. X4: Snažím se nalézt co nejlepší nabídku. X5: Nemám o nákupy velký zájem. X6: Srovnávání cen umožňuje úsporu mnoha peněz.

V tabulce 1a jsou uvedena vstupní data.

Použijeme-li **eukleidovskou vzdálenost** jednotek (respondentů) a při shlukování jako kritérium **nejmenší vzdálenosti** (metoda nejbližšího souseda - nearest neighbor) mezi nimi, získáme následující řešení po použití například programu Statgraphics 5.0Plus (následuje výstup programu v angličtině, je doplněn českým komentářem autora).

Tab. 1a: Vstupní data

Zákazník	X1	X2	X3	X4	X5	X6
1	4	6	3	7	2	7
2	2	3	2	4	7	2
3	7	2	6	4	3	3
4	4	6	4	5	3	6
5	1	3	2	2	6	4
6	6	4	6	3	3	4
7	5	3	6	3	3	4
8	7	3	7	4	1	4
9	2	4	3	3	6	3
10	3	5	3	6	4	6
11	1	3	2	4	5	3
12	5	4	4	4	2	3
13	2	2	1	5	4	4
14	4	6	4	6	4	7
15	6	5	4	2	1	4
16	3	4	4	6	4	7
17	4	4	7	2	2	5
18	3	6	2	6	4	3
19	5	4	7	3	2	3
20	2	3	1	4	5	4

Zdroj: vlastní

Cluster Analysis

Analysis Summary:

- Data variables: X1, X2, X3, X4, X5, X6
- Number of complete cases: 20
- Clustering Method: Nearest Neighbor (Single Linkage)
- Distance Metric: Euclidean
- Clustering Method: Nearest Neighbor (Single Linkage)
- Distance Metric: Euclidean

V tab. 1b jsou jednotliví respondenti (Row 1-20) přiřazeni do 3 shluků (Cluster 1-3). Pro interpretaci vytvořených tří shluků bude užitečné, spočteme-li ještě pro všech šest proměnných skupinové průměry (tab. 1c).

Nejvyšší průměrné hodnoty jsme označili hvězdičkou. V první skupině (Shluk 1) tedy zjišťujeme nejvýraznější souhlas s výroky X4 a X6; jedná se tedy o osoby preferující cenu. Ve druhé skupině

Tab. 1b: Membership Table

Row	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Cluster	1	2	3	1	2	3	3	3	2	1	2	3	2	1	3	1	3	1	3	2

Zdroj: vlastní

Tab. 1c: Skupinové průměry

Cluster	X1	X2	X3	X4	X5	X6
1	3,5	5,5	3,33333	6,0*	3,5	6,0*
2	1,66667	3,0	1,83333	3,66667	5,5*	3,33333
3	5,625*	3,625	5,875*	3,125	2,125	3,75

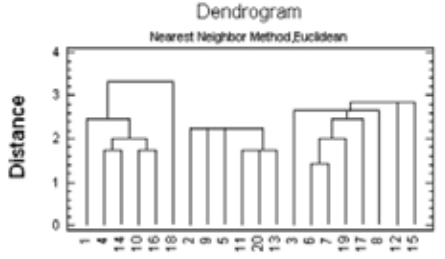
Zdroj: vlastní

(Shluk 2) jsou reakce na výroky X1 a X3 nesouhlasné a naopak je nejvýraznější souhlas s výrokem X5 (nezajímám se o nakupování) - jedná se zákazníky, kteří neradi nakupují (nejspíše muži). Do poslední skupiny (Shluk 3) jsou pak zřejmě zařazeny osoby, pro které nakupování baví (zřejmě ženy), (souhlas s X1 a s X3).

Graficky potom danou situaci znázorňuje tzv. **dendrogram** (obr. 1) a také další doplňující grafy (grafické výstupy z programu Statgraphics 5.0), (obr. 2). Na prvním grafu lze identifikovat následující tři shluky:

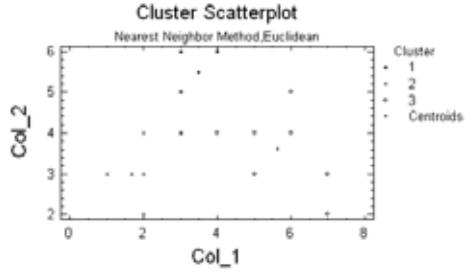
Shluk 1: 1, 4, 14, 10, 16, 18.
Shluk 2: 2, 9, 5, 11, 20, 13.
Shluk 3: 3, 6, 7, 19, 17, 8, 12, 15.

Obr. 1: Dendrogram



Zdroj: vlastní zpracování v programu Statgraphics 5.0

Obr. 2: Bodový graf



Zdroj: vlastní zpracování v programu Statgraphics 5.0

Příklad 2 - dvourozměrná shluková analýza

K ilustraci dvourozměrného shlukování použijeme data z následující tabulky. Jedním sledovaným znakem je název výrobku (symbolicky I-VI), druhým je nejvyšší dosažený stupeň vzdělání (6 úrovní P_1-P_6). Tabulka 2a obsahuje pro každou kombinaci kódů průměrné hodnocení výrobku (ze známek 1-5). V tabulkách 2b a 2c jsou vypočte-

Tab. 2a: Vstupní data

Stupeň vzdělání	Výrobek					
	I	II	III	IV	V	VI
P_1 základní škola	4,4	1,3	2,4	4,3	2,4	1,2
P_2 střední škola bez maturity	4,2	2,1	3,5	4,4	3,4	1,8
P_3 střední škola s maturitou	1,2	3,5	4,3	2,3	3,2	3,6
P_4 vyšší odborná škola	2,3	3,6	4,2	1,5	3,1	3,7
P_5 VŠ bakalářské	3,6	4,2	1,7	3,5	2,8	3,5
P_6 VŠ magisterské	3,6	4,3	2,8	2,6	1,9	3,4

Zdroj: Vlastní

Tab. 2b: Základní charakteristiky proměnných P_1-P_6

Stupeň vzdělání	Průměr	Sm. odchylka
P_1	2,666667	1,402379
P_2	3,233333	1,070825
P_3	3,016667	1,101665
P_4	3,066667	1,001332
P_5	3,216667	0,865833
P_6	3,100000	0,843801

Zdroj: Vlastní

ny základní charakteristiky pro stupeň vzdělání a výrobek.

Uvedený typ analýzy umožňuje například programový systém STATISTICA 6.0Cz. Jako výsledky lze získat barevný graf bloků, průměry a směrodatné odchylky pro jednotlivé sloupce a řádky. Stupnice barev v programu v grafu začíná tmavě zelenou, pokračuje přes světlejší odstíny až k červené. Díky nevhodné úpravě pro tisk zde není uveden. Obdobné výsledky jako u grafu získáme v tabulce 2d.

V tabulce 2d je změněno pořadí výrobků tak, aby podobné spolu sousedily. Vzhledem k tomuto přeuspořádaní je zřejmé, že byly podobně hodnoceny výrobky I a VI, IV a III, V a II. Dále je možné z tabulky 2d zjistit, že v hodnocení se shodovali respondenti se základním vzděláním a se středním vzděláním bez maturity (P_1 a P_2), se středním s maturitou a vyšším odborným (P_3 a P_4) a nakonec s bakalářským a magisterským (P_5 a P_6).

Tab. 2c: Základní charakteristiky proměnných I-VI

Výrobek	Průměr	Sm. odchylka
I	3,216667	1,230312
II	3,166667	1,206096
III	3,150000	1,032957
IV	3,100000	1,161034
V	2,800000	0,562139
VI	2,866667	1,080123

Zdroj: Vlastní

Tab. 2d: Přeskupená matice dat

	I	VI	IV	III	V	II
P_1	4,4	1,2	4,3	2,4	2,4	1,3
P_2	4,2	1,8	4,4	3,5	3,4	2,1
P_3	1,2	3,6	2,3	4,3	3,2	3,5
P_4	2,3	3,7	1,5	4,2	3,1	3,6
P_5	3,6	3,5	3,5	1,7	2,8	4,2
P_6	3,6	3,4	2,6	2,8	1,9	4,3

Zdroj: Vlastní

6. Závěr

Jednou z příležitostí v procesu dobývání znalostí je použití shlukové analýzy. Shluková analýza patří v data miningu k metodám bez učitele. Oblasti možných aplikací jsou značně široké. Jmenujme například aplikace ekonomické, lékařské, zemědělské, chemické nebo astronomické. Základní principy a metody shlukové analýzy byly popsány v tomto článku. Byly dále demonstrovány na dvou praktických příkladech z oblasti marketingových výzkumů za použití statistických výpočetních prostředků (Statgraphics 5.0 Plus a Statistica 6.0Cz). V marketingových výzkumech se jeví použití shlukové analýzy jako velmi užitečné.

Literatura:

[1] KLÍMEK, P. *Získávání znalostí z podnikových dat (data mining)*. Disertační práce. Zlín: UTB, FaME, 2003.

[2] KLÍMEK, P. Data mining a jeho využití. *E+M Ekonomie a Management,* 2005, roč. 8, č. 3, s. 128 - 135. ISSN 1212-3609.

[3] KLÍMEK, P. Stručná zpráva o data miningu v České republice. I*nformační Bulletin ČStS*, 2006, roč. 17, č. 2 - 3, s. 6 - 11. ISSN 1210-8022.

[4] NOVÁK, V. *Základy fuzzy modelování*. Praha: Ben - technická literatura, 2002. ISBN 80-7300-009-1.
[5] PARR RUD, O. *Data mining (Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. Praha: Computerpress, 2002. ISBN 80-7226-577-6.
[6] ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V. *Shluková analýza dat*. Praha: Professional Publishing, 2007. ISBN 978-80-86946-26-9.
[7] ŽAMBOCHOVÁ, M. Data mining methods with trees. *E+M Ekonomie a Management*, 2008, roč. 11, č. 1. ISSN 1212-3609.
[8] ŽAMBOCHOVÁ, M., Použití stromů ve statistice. *Sborník Ekonomika, regiony a jejich výhledy*. Ústí nad Labem: Univerzita J.E.Purkyně v Ústí nad Labem, 2006. ISBN 80-7044-795-8.

Ing. Petr Klímek, Ph.D.
Univerzita Tomáše Bati ve Zlíně
Fakulta managementu a ekonomie
Ústav informatiky a statistiky
klimek@fame.utb.cz

Doručeno redakci: 18. 10. 2007
Recenzováno: 10. 12. 2007
Schváleno k publikování: 7. 4. 2008

ABSTRACT

DATA MINING WITH CLUSTERING

Petr Klímek

Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. It is concerned with the secondary analysis of large databases in order to find previously unsuspected relationships which are of interest or value to the database owners. There are two keys to success in data mining. First is coming up with a precise formulation of the problem you are trying to solve. A focused statement usually results in the best payoff. The second key is using the right data. After choosing from the data available to you, or perhaps buying external data, you may need to transform and combine it in significant ways. New problems arise, partly as a consequence of the sheer size of the data sets involved, and partly because of issues of pattern matching. However, since statistics provides the intellectual glue underlying the effort, it is important for statisticians to become involved. There are very real opportunities for statisticians to make significant contributions.

The main definition of data mining and the special data mining tasks are mentioned in the first part of this paper. The data mining problem was also discussed in previous issues of E+M. One method (clustering) was chosen to be a subject of this article.

One of the opportunities to gain knowledge from data is a use of clustering analysis. Clustering analysis belongs to unsupervised methods of data mining. We put here a focus on this method. Some basic principles are described in the second part of this paper. This method is examined on two examples from the marketing field. In the first example is used software Statgraphics 5.0Plus (www.statgraphics.com) to solve clustering problem (nearest neighbour algorithm and Eucleidian distance), and in the second example is used Statistica 6.0Cz software (from Statoft, Inc., www.statsoft.com or www.statsoft.cz).

But the building models is only one step in knowledge discovery. It is vital to properly collect and prepare the data, and to check the models against the real world. The „best" model is often found after building models of several different types, or by trying different technologies or algorithms.

Key Words: *data mining, clustering, nearest neighbour method, dendrogram*

JEL Classification: *C1, C19*