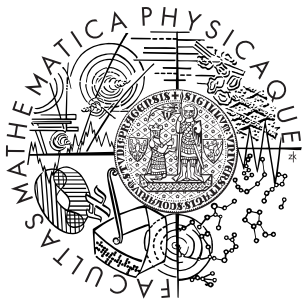
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Petr Strejc

Shluková analýza ve financích

Katedra pravděpodobnosti a matematické statistiky

Doc. RNDr. Jan Hurt, CSc.

Studijní program: MATEMATIKA, obor: Finanční matematika

2009

Rád bych poděkoval všem, kteří mě při vytváření této práce podporovali, jmenovitě pak Josefu Rubášovi za pomoc s LaTeXem a především svému vedoucímu Doc. RNDr. Janu Hurtovi, CSc. nejen za pomoc a cenné rady.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 3.8.2009 Petr Strejc

Obsah

Název práce: Shluková analýzy ve financích
Autor: Petr Strejc
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Doc. RNDr. Jan Hurt, CSc.
e-mail vedoucího: hurt@karlin.mff.cuni.cz

Abstrakt: Cílem této práce je seznámit zájemce se základními myšlenkami a principy shlukové analýzy a demonstrovat její použití na datech z pražské a newyorské burzy. Pro porozumění tomuto textu není nutná dřívější znalost problematiky shlukování, neboť práce čtenáře do celé situace nejprve uvede. Požadavkem je pouze základní znalost matematiky a matematické terminologie. Po úvodu následují definice a vysvětlení základních pojmů, na kterých jsou postaveny algoritmy shlukové analýzy. Dva z nich jsou v práci podrobně popsány a doplněny krátkými ilustrativními příklady. Aplikace shlukové analýzy na společnostech z pražské a newyorské burzy je ukázkou shlukování za použití matematického software Wolfram Mathematica a zároveň slouží jako příklad využití shlukové analýzy ve financích.

Klíčová slova: shluková analýza, shlukování, finance, Mathematica

Title: Cluster analysis in finance
Author: Petr Strejc
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. RNDr. Jan Hurt, CSc.
Supervisor's e-mail address: hurt@karlin.mff.cuni.cz

Abstract: The aim of the present thesis is to present basic ideas and principles of the cluster analysis and demonstrate them on real data from Prague and New York Stock Exchange. To understand the text only the elements of mathematics is necessary. After the introduction the definitions and basic concepts used in algorithms are given. Two of them are described in details and illustrated by examples. Application of the cluster analysis is realized using Wolfram's Mathematica software. It also shows the usefulness of this method in finance.

Keywords: cluster analysis, clustering, finance, Mathematica

Kapitola 1

Úvod

Název shluková analýza (přesněji anglický ekvivalent Cluster Analysis) poprvé použil R.C. Tryon již před 70-ti lety. Nejstarší definice formulovaná právě Tryonem popisuje shlukovou analýzu jako „obecně logický postup formulovaný jako procedura, pomocí níž objektivně seskupujeme jedince do skupin na základě jejich podobností a rozdílností“.

Ačkoliv si to ani nemusíme uvědomovat, je pro nás vytváření shluků a práce s nimi naprosto běžná. Například na uklízení můžeme nahlížet jako na proces, ve kterém se snažíme rozdělit věci v domácnosti do pokud možno homogenních skupin, abychom si usnadnili jejich používání, vyhledávání a skladování. Vlastnosti, ke kterým budeme v takovém případě přihlížet, mohou být podobnost účelu použití, četnost používání nebo třeba i rozměry.

Zboží v obchodech je pro lepší orientaci nakupujících také tříděno do skupin. Očekáváme, že podobné zboží se bude nacházet poblíž. Podobností v tomto případě můžou být požadavky výrobku na skladování (např. zmrzlé výrobky v mrazících boxech), složení (např. mléčné výrobky) nebo způsob použití. Z tohoto příkladu je patrné, že některé z vlastností, na základě kterých rozdělení do shluků vytváříme, mohou být navzájem závislé (např. složení a požadovaný způsob skladování). Tato skutečnost nic nekomplikuje, neboť ve shlukové analýze nezávislost není obecně požadována.

Motivace pro shlukování dat do skupin je tedy zřejmá: členění do skupin nám usnadňuje práci s rozsáhlým souborem dat, usnadňuje orientaci a na základě takovéhoto rozdělení můžeme být schopni odhalit některé neznáme souvislosti. Ne všechny úlohy vyžadující rozdělení do shluků můžeme řešit intuitivně jako třeba při uklízení. Právě v takových situacích postupujeme podle určitého algoritmu a na základě objektivních kritérií.

Shluková analýza netvoří ucelenou teorii, ale zastřešuje celou řadu metod se společným cílem, ale rozdílnými postupy. Nejčastěji je používána v situacích, kdy nemáme a priori žádnou hypotézu nebo znalost o struktuře dat. Shluková analýza je tedy velmi často používána například pro data mining. Díky vysoké poptávce po zjišťování informací z dat prožívají dnes metody shlukové analýzy své období renesance.

Široká uplatnitelnost je hlavním z důvodů, proč je shluková analýza jednou z nej-

používanějších vícerozměrných statistických metod. Má však i svá úskalí, a sice ve vhodné volbě proměnných, na základě kterých data rozdělujeme, v použité míře podobnosti resp. nepodobnosti a metodě výpočtu a v neposlední řadě v interpretaci výsledků. Je zřejmé, že dobrých výsledků dosáhneme zejména tam, kde objekty mají tendenci se přirozeně seskupovat do shluků.

Kapitola 2

Formulace úlohy

Máme n objektů, kde každý objekt je charakterizován p znaky.

$$\begin{aligned} C &= \{1, 2, 3, \ldots, n\} && \text{je množina všech objektů.} \\ \boldsymbol{X}_i &= (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})^\top && \text{jsou p-členné vektory pozorování.} \\ \boldsymbol{X} &= (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots, \boldsymbol{X}_n) && \text{je n-rozměrný vektor znaků objektů (matice).} \end{aligned}$$

Našim cílem je vytvořit disjunktní rozklad do jednotlivých shluků tak, aby si objekty uvnitř jedné skupiny byly co nejvíce podobné a objekty patřící do různých skupin si byly podobné co nejméně. Vytváříme tedy k shluků $C_1, C_2, C_3, \ldots, C_k$ takových, že

$$\begin{gathered} C_i \text{ je neprázdná podmnožina } C \quad \forall\, i \in \{1, 2, 3, ..., k\}, \\ C_i \cap C_j = \emptyset \quad \text{pro } i \neq j, \quad i, j \in \{1, 2, 3, ..., k\}, \\ C_1 \cup C_2 \cup \cdots \cup C_k = C. \end{gathered}$$

Teoreticky je možné, aby počet shluků k dosáhl počtu objektů n, praktický význam však pro nás bude mít pouze takový počet shluků, který je výrazně nižší než je teoretické maximum.

2.1 Rozdělení metod

Vzhledem k faktu, že různých shlukovacích metod je nemalé množství, rozdělujeme metody shlukové analýzy ne na základě použitých algoritmů, ale podle cílů, k nimž směřují, na hierarchické a nehierarchické.

Nehierarchické úlohy: hledáme disjunktní rozklad $C = C_1 \cup C_2 \cup \cdots \cup C_k$ při známém anebo neznámém k.

Hierarchické úlohy: vytváříme hierarchický strom pomocí posloupnosti rozkladu $S^{(1)}, S^{(2)}, \ldots, S^{(k)}$, viz [3].

V případě *aglomerativních* metod začínáme s jednotlivými objekty a nejpodobnější z nich spojujeme do skupin, se kterými dále nakládáme jako se samostatnými objekty. Algoritmus končí v okamžiku, kdy máme jedinou skupinu:

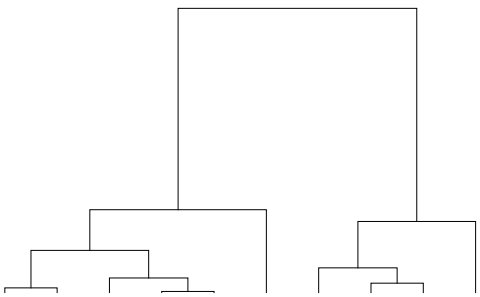
$$S^{(0)} = \{\{1\}, \{2\}, \{3\}, \ldots, \{n\}\}$$
$$S^{(t)} = C$$
$$\forall\ C^* \in S^{(i+1)} \quad \text{existují } C_1, C_2, ..., C_m \in S^{(i)} : \ C^* = C_1 \cup C_2 \cup \cdots \cup C_m.$$

U *divizivních* metod začínáme z "opačného konce". Celý soubor objektů dělíme do několika (obvykle do dvou) skupin. Na nově vzniklé skupiny aplikujeme stejný algoritmus až do okamžiku, kdy nám vznikne n skupin:

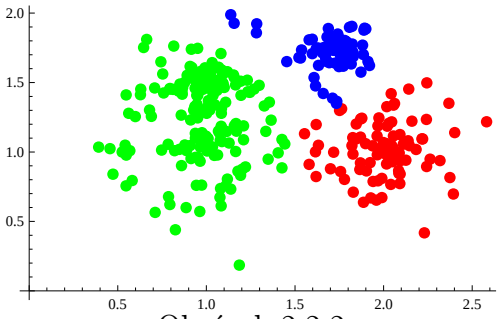
$$S^{(0)} = C$$
$$S^{(t)} = \{\{1\}, \{2\}, \{3\}, \ldots, \{n\}\}$$
$$\forall\ C^* \in S^{(i)} \quad \text{existují } C_1, C_2, ..., C_m \in S^{(i+1)} : \ C^* = C_1 \cup C_2 \cup \cdots \cup C_m.$$

2.2 Výstup shlukové analýzy

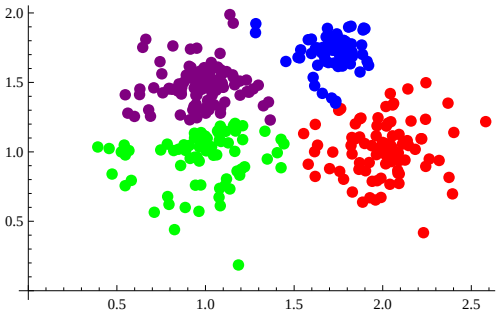
Výstup shlukové analýzy může být číselný nebo grafický. V případě číselného výstupu přiřadíme každému objektu číslo shluku, do kterého byl zařazen. Pro zobrazení výsledků některé hierarchické metody obvykle používáme dendrogram - dvojrozměrný graf, ve kterém jedna osa představuje ekvidistantně rozdělené objekty a druhá osa zobrazuje vzdálenosti spojovaných shluků (viz obrázek 2.2.1). U nehierarchických metod je na rozdíl od grafického znázornění označení číslem příslušného shluku použitelné vždy, grafický výstup na druhou stranu může být názornější a přehlednější. Obrázky 2.2.2, 2.2.3 a 2.2.4 znázorňují 3, 4 a 5 shluků vytvořených ze stejných dvojrozměrných dat (Zdroj: [10]).



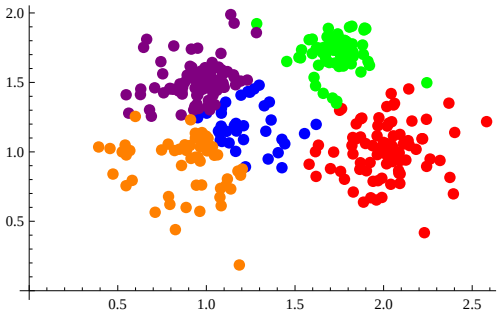
Obrázek 2.2.1



Obrázek 2.2.2



Obrázek 2.2.3



Obrázek 2.2.4

2.3 Typy proměnných a jejich úprava

2.3.1 Kvantitativní znaky

Kvantitativní znaky vyjadřují množství nebo jinou veličinu, která nabývá obvykle reálných hodnot nebo hodnot nějaké podmnožiny $\mathbb{R}$ (například celá čísla). Vzhledem ke skutečnosti, že samotná volba jednotek pro složky vektoru $\boldsymbol{X}_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})^\top$ může výrazně ovlivnit výslednou vzdálenost objektů, bývá užitečné nejprve data standardizovat. Jedním z možných postupů je nejprve spočítat průměrnou hodnotu j-tého znaku

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

a rozptyl

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2.$$

Standardizované hodnoty z_{ij} získáme transformací

$$z_{ij} = \frac{x_{ij} - x_j}{s_j}.$$

2.3.2 Binární znaky

Pomocí binárních znaků rozlišujeme, do jaké ze dvou skupin objekt patří. Veličina x_{ij} tedy nabývá hodnot 0 nebo 1 podle toho, zda daný objekt má nebo nemá požadovanou vlastnost.

2.3.3 Kvalitativní znaky

Kvalitativní znaky nabývají hodnot nějaké konečné množiny T. Takovýto znak můžeme převést na kvantitativní nebo binární.

V případě, že hodnoty T *nelze uspořádat* nějakým "rozumným" způsobem, můžeme takovou proměnnou vyjádřit pomocí binárních proměnných. Takový případ může nastat v situaci, kdy rozlišujeme 3 barvy (červená, zelená a modrá), a uspořádání podle vlnové délky apod. by pro náš experiment nemělo smysl.

Původní proměnnou nabývající hodnot "červená", "zelená" nebo "modrá" nahradíme třemi binárními proměnnými označenými č,z,m způsobem znázorněným v tabulce níže.

	č	**z**	**m**
červená	1	0	0
zelená	0	1	0
modrá	0	0	1

Pokud hodnoty T *lze uspořádat*, můžeme takovou proměnnou opět nahradit několika binárními anebo jednou kvantitativní proměnnou. Ilustrativním příkladem je dosažené vzdělání (základní, středoškolské, vysokoškolské). Nahrazením binárními proměnnými (SS, VS) získáme:

	SS	**VS**
základní	0	0
středoškolské	1	0
vysokoškolské	1	1

Kvantitativní proměnná nahrazující slovní vyjádření dosaženého vzdělání může např. nabývat hodnot 1 pro základní, 2 pro středoškolské a 3 pro vysokoškolské vzdělání.

Kapitola 3

Vzdálenosti mezi objekty

Existují dva základní přístupy, jak vzdálenost mezi objekty definovat.

3.1 Míra podobnosti

Obvykle označovaná jako s (similarity), je zobrazení $\boldsymbol{X} \times \boldsymbol{X} \longrightarrow [0, \infty)$ splňující tyto požadavky:

$$\begin{aligned} s(\boldsymbol{X}_i, \boldsymbol{X}_j) &= s(\boldsymbol{X}_j, \boldsymbol{X}_i) \qquad \forall \boldsymbol{X}_i, \boldsymbol{X}_j \in \boldsymbol{X} \qquad \text{(symetrie)}, \\ s(\boldsymbol{X}_i, \boldsymbol{X}_i) &= \sup_{\boldsymbol{X}_j \in \boldsymbol{X}} s(\boldsymbol{X}_i, \boldsymbol{X}_j) \qquad \forall \boldsymbol{X}_i \in \boldsymbol{X}. \end{aligned}$$

Míra podobnosti s nabývá tím vyšší hodnoty, čím jsou si objekty navzájem podobnější. Při určování míry podobnosti dvou stejných objektů se můžeme dostat do nesnází, protože určení supréma nemusí být vždy jednoduché. Proto ve shlukové analýze obvykle místo podobnosti dvou objektů určujeme jejich nepodobnost.

3.2 Míra nepodobnosti

Pro označení míry nepodobnosti používáme obvykle písmeno d (dissimilarity, distance). Stejně jako v případě míry podobnosti je d zobrazení $\boldsymbol{X} \times \boldsymbol{X} \longrightarrow [0, \infty)$ a splňuje:

$$\begin{aligned} d(\boldsymbol{X}_i, \boldsymbol{X}_j) &= d(\boldsymbol{X}_j, \boldsymbol{X}_i) \qquad \forall \boldsymbol{X}_i, \boldsymbol{X}_j \in \boldsymbol{X} \qquad \text{(symetrie)}, \\ d(\boldsymbol{X}_i, \boldsymbol{X}_i) &= 0 \qquad \forall \boldsymbol{X}_i \in \boldsymbol{X}. \end{aligned}$$

V praxi se jako míra nepodobnosti často používá metrika na $\mathbb{R}^n$, která oproti míře nepodobnosti navíc ještě splňuje:

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0 \Longleftrightarrow \boldsymbol{X}_i = \boldsymbol{X}_j,$$
$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) + d(\boldsymbol{X}_j, \boldsymbol{X}_k) \geq d(\boldsymbol{X}_i, \boldsymbol{X}_k).$$

Euklidovská vzdálenost

Jde o pravděpodobně nejčastěji používaný výpočet vzdálenosti založený na geometrické vzdálenosti ve vícedimenzionálním prostoru $\mathbb{R}^n$. Stejně jako u jiných měr nepodobnosti jsme schopni napočítat symetrickou matici $\boldsymbol{D}$ typu $n \times n$ s nulovými hodnotami na hlavní diagonále, kde prvek na pozici i, j nabývá hodnoty

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}.$$

Nevýhodou tohoto typu vzdálenosti je skutečnost, že výsledek je silně ovlivněn zvoleným měřítkem jednotlivých znaků objektů. Pokud například u jedné proměnné místo metrů použijeme vyjádření v centimetrech, výsledná vzdálenost objektů se může velmi změnit, aniž by ovšem došlo k faktické změně vstupních dat. Proto je vhodné data před samotným výpočtem vzdálenosti nejprve upravit tak, aby poměry rozdílů jednotlivých proměnných odpovídaly jejich významnosti. Jedním z možných řešení je nejprve vstupní data upravit tak, aby měla nulovou střední hodnotu a jednotkový rozptyl (viz výše) a případně místo klasické euklidovské vzdálenosti použít euklidovskou vzdálenost s vhodně zvolenými vahami.

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2},$$

kde w_k je váha k-tého znaku.

Druhá mocnina euklidovské vzdálenosti

Využívaná v situacích, kdy požadujeme progresivnější výpočet vzdálenosti v porovnáním s klasickou euklidovskou vzdáleností. Vyšší hodnoty tedy mají větší váhu než hodnoty nízké.

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

Hemmingova vzdálenost

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Čebyševova vzdálenost

Tuto metriku využijeme zejména v případě, kdy dva objekty považujeme za vzdálené právě tehdy, když existuje proměnná, ve které jsou významně rozdílné:

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \max_{k \in \{1,2,\ldots,p\}} |x_{ik} - x_{jk}| .$$

Mocninná vzdálenost

Mocninná vzdálenost nám umožňuje vhodnou volbou parametrů r a q navolit významnost "vzdálenosti" znaků dvou objektů:

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt[r]{\sum_{k=1}^{p} \left(|x_{ik} - x_{jk}|\right)^q}.$$

Předchozí metriky jsou tedy speciálními případy mocninné vzdálenosti.
Euklidovská vzdálenost ($r = q = 2$).
Kvadrát euklidovské vzdálenosti ($r = 1,\ q = 2$).
Hemmingova vzdálenost ($r = q = 1$).
Čebyševova vzdálenost ($r = q \longrightarrow \infty$).
Minkovského vzdálenost ($r = q$).

Mahalanobisova vzdálenost

Velkou výhodou Mahalabisovy vzdálenosti je fakt, že bere v úvahu rozdílnou variabilitu dat a jejich vzájemnou korelaci.

Alternativní způsob pro výpočet euklidovské vzdálenosti je:

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = ((\boldsymbol{X}_i - \boldsymbol{X}_j)^\top \ \boldsymbol{I} \ (\boldsymbol{X}_i - \boldsymbol{X}_j))^{1/2},$$

kde $\boldsymbol{I}$ je jednotková diagonální matice. Pokud místo matice $\boldsymbol{I}$ použijeme $\boldsymbol{V}^{-1}$, kde $\boldsymbol{V}$ je výběrovou varianční maticí

$$\boldsymbol{V} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right) \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^\top,$$

získáme vzdálenost (Mahalanobisovu) s vahami závisejícími na rozptylu a korelaci dat

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = ((\boldsymbol{X}_i - \boldsymbol{X}_j)^\top \ \boldsymbol{V}^{-1} \ (\boldsymbol{X}_i - \boldsymbol{X}_j))^{1/2}.$$

Koeficient asociace

$$s(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{\sum_{k=1}^{p} I_{ijk}}{p},$$

kde $I_{ijk} = 1 \iff x_{ik} = x_{jk}$ a $I_{ijk} = 0$ v jiných případech. Koeficient asociace, nazývaný také koeficient souhlasu, je použitelný především pro binární nebo kvalitativní znaky, které nelze seřadit. V případě binárních znaků můžeme koeficient souhlasu zapsat alternativně jako

$$s(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{\sum_{k=1}^{p} \left[x_{ik} \cdot x_{ik} + (1 - x_{ik})\,(1 - x_{jk})\right]}{p}.$$

Je zřejmé, že výraz uvnitř sumy nabývá hodnoty 1 právě tehdy, když $x_{ik} = x_{jk}$ a jinak je roven nule. Výše definované koeficienty souhlasu nabývají hodnot z intervalu $\langle 0, 1\rangle$, kde 1 znamená naprostou shodu a 0 maximální odlišnost. Můžeme tedy definovat koeficient nesouhlasu (vzdálenost) d,

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 1 - s(\boldsymbol{X}_i, \boldsymbol{X}_j).$$

Další míry nepodobnosti jsou popsány např. v [7].

Kapitola 4

Vzdálenosti mezi shluky

Pokud každý objekt představuje jeden jednoprvkový shluk, je vzdálenost mezi shluky jednoduše definovaná jako vzdálenost mezi objekty. Když však spojíme několik objektů do shluků, budeme potřebovat určit vzdálenost shluků mezi sebou navzájem. To budeme potřebovat později, abychom byli schopni určit, jestli jsou si dva shluky dostatečně podobné, abychom je mohli sloučit do jednoho, nebo ne.

Nechť d je libovolná míra nepodobnosti mezi objekty, A a B jsou libovolné shluky. Míra nepodobnosti shluků δ musí splňovat:

$$\begin{aligned} \delta(A, A) &= 0 \\ \delta(A, B) &\geq 0 \\ \delta(A, B) &= \delta(B, A). \end{aligned}$$

Mezi nejčastěji používané míry nepodobnosti shluků definované pomocí některé míry nepodobnosti objektů patří:

Princip nejbližšího souseda

$$\delta(A, B) = \min_{a \in A \ b \in B} d(a, b)$$

Princip nejvzdálenějšího souseda

$$\begin{aligned} \delta(A, B) &= \max_{a \in A \ b \in B} d(a, b) \\ \delta(A, A) &= 0 \end{aligned}$$

Princip průměrné vzdálenosti prvků

$$\delta(A,B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a,b),$$

kde $|A|$ je počet objektů ve shluku A.

Centroidní metoda (princip vzdálenosti průměrů)

$$\delta(A,B) = d(t_a, t_b),$$

kde t_a a t_b jsou těžiště shluků A a B.

$$\begin{aligned} t_a &= \frac{1}{|A|} \sum_{a \in A} a \\ t_b &= \frac{1}{|B|} \sum_{b \in B} b \end{aligned}$$

Kolmogorovova zobecněná vzdálenost

$$\delta_r(A,B) = \left(\frac{\sum_{a \in A} \sum_{b \in B} d^r(a,b)}{|A| \cdot |B|} \right)^{1/r}.$$

První tři výše popsané míry jsou tedy speciálními případy zobecněné Kolmogorovovy vzdálenosti.

$r \longrightarrow -\infty$ odpovídá principu nejbližšího souseda,
$r \longrightarrow +\infty$ odpovídá principu nejvzdálenějšího souseda,
$r = 1$ odpovídá principu průměrné vzdálenosti.

Mahalanobisova zobecněná vzdálenost

$$\delta(A,B) = \left((t_a - t_b)^\top \boldsymbol{V}^{-1} (t_a - t_b) \right)^{1/2},$$

kde $\boldsymbol{V}$ je výběrová varianční matice těžišť všech shluků.

Kapitola 5

Ukázky shlukovacích algoritmů

5.1 Hierarchické shlukování

5.1.1 Metoda nejbližšího souseda

Tato jednoduchá aglomerativní metoda podrobně popisovaná v [3] spojuje ty shluky, jejichž některé dva prvky jsou si nejblíže. Vytváří tedy obvykle shluky ve tvaru řetězce a s ohledem na tuto skutečnost by tato metoda měla být také používána. Pokud je zřetězování pro naše data nevhodné, použijeme jinou metodu.

Při tvorbě hierarchického stromu pomocí metody nejbližšího souseda si nejprve spočítáme matici míry nepodobnosti mezi prvky – matici $\boldsymbol{D}$. $\boldsymbol{D}$ je tedy matice typu $n \times n$, která má na místě i, j prvek $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$, kde d je námi vybraná míra nepodobnosti. Vzhledem k symetrii matice a nulovosti prvků na hlavní diagonále, stačí uvažovat jen prvky pod hlavní diagonálou, které seřadíme podle velikosti do neklesající posloupnosti

$$d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{\binom{n}{2}}.$$

První rozklad $S^{(0)}$je rozkladem na jednoprvkové shluky.

$$S^{(0)} \quad = \quad \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$$

Následující sled kroků opakujeme, dokud nevytvoříme jediný shluk.

V k-tém kroku ($1 \leq k \leq \binom{n}{2}$) použijeme prvek posloupnosti $d_{(k)} = d(\boldsymbol{X}_i, \boldsymbol{X}_j)$. Pokud už objekty $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ patří do stejného shluku, posuneme se ke kroku $k + 1$. Pokud $\boldsymbol{X}_i$ a $\boldsymbol{X}_j$ patří do dvou různých shluků, tyto shluky spojíme, protože jsou si nejblíže (ve smyslu nejbližšího souseda).

5.1.2 Příklad

Máme 6 jednorozměrných objektů, jako míru nepodobnosti uvažujeme euklidovskou metriku.

$$\begin{aligned} \boldsymbol{X} &= (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5, \boldsymbol{X}_6) \\ \boldsymbol{X} &= (2 \quad 2,4 \quad 6,8 \quad 3,4 \quad 5,2 \quad 7,7) \end{aligned}$$

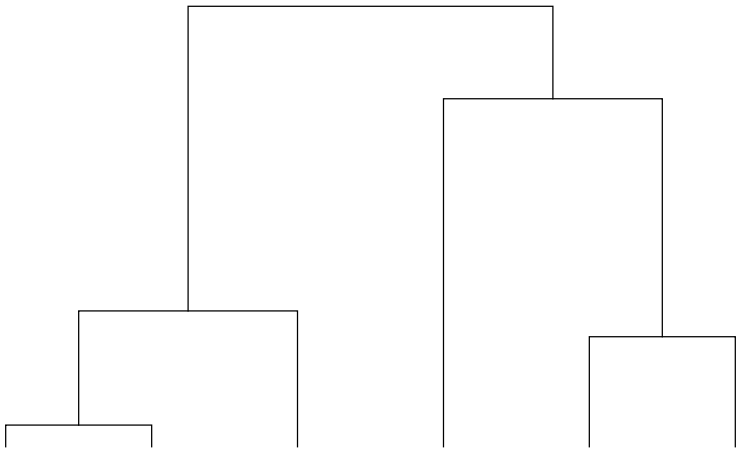
Matice $\boldsymbol{D}$ má tvar:

$$\boldsymbol{D} = \begin{pmatrix} 0,0 & 0,4 & 4,8 & 1,4 & 3,2 & 5,7 \\ 0,4 & 0,0 & 4,4 & 1,0 & 2,8 & 5,3 \\ 4,8 & 4,4 & 0,0 & 3,4 & 1,6 & 0,9 \\ 1,4 & 1,0 & 3,4 & 0,0 & 1,8 & 4,3 \\ 3,2 & 2,8 & 1,6 & 1,8 & 0,0 & 2,5 \\ 5,7 & 5,3 & 0,9 & 4,3 & 2,5 & 0,0 \end{pmatrix}$$

Jednotlivé kroky, ve kterých vytváříme nové shluky, jsou pro přehlednost zobrazeny v tabulce 5.1.2. Výsledný dendrogram je zachycený na obrázku 5.1.2.

k	$d_{(k)}$	shluky ($\mathbf{X}_1$ nahrazeno 1 atd.)
1	$d(\mathbf{X}_2, \boldsymbol{X}_1) = 0,4$	{{1,2}, 3 ,4 , 5 , 6}
2	$d(\mathbf{X}_6, \boldsymbol{X}_3) = 0,9$	{{1, 2}, {3, 6} ,4 , 5}
3	$d(\mathbf{X}_4, \boldsymbol{X}_2) = 1,0$	{{1, 2, 4}, {3, 6}, 5}
4	$d(\mathbf{X}_4, \boldsymbol{X}_1) = 1,4$	{{1, 2, 4}, {3, 6}, 5}
5	$d(\mathbf{X}_5, \boldsymbol{X}_3) = 1,6$	{{1, 2, 4}, {3, 5, 6}}
6	$d(\mathbf{X}_5, \boldsymbol{X}_4) = 1,8$	{{1, 2, 4}, {3, 5, 6}}
7	$d(\mathbf{X}_6, \boldsymbol{X}_5) = 2,5$	{{1, 2, 4}, {3, 5, 6}}
8	$d(\mathbf{X}_5, \boldsymbol{X}_2) = 2,8$	{{1, 2, 3, 4, 5, 6}}
9	$d(\mathbf{X}_5, \boldsymbol{X}_1) = 3,2$	{{1, 2, 3, 4, 5, 6}}
10	$d(\mathbf{X}_4, \boldsymbol{X}_3) = 3,4$	{{1, 2, 3, 4, 5, 6}}
11	$d(\mathbf{X}_6, \boldsymbol{X}_4) = 4,3$	{{1, 2, 3, 4, 5, 6}}
12	$d(\mathbf{X}_3, \boldsymbol{X}_2) = 4,4$	{{1, 2, 3, 4, 5, 6}}
13	$d(\mathbf{X}_3, \boldsymbol{X}_1) = 4,8$	{{1, 2, 3, 4, 5, 6}}
14	$d(\mathbf{X}_6, \boldsymbol{X}_2) = 5,3$	{{1, 2, 3, 4, 5, 6}}
15	$d(\mathbf{X}_6, \boldsymbol{X}_1) = 5,7$	{{1, 2, 3, 4, 5, 6}}

Tabulka 5.1.2



Obrázek 5.1.2

5.2 Nehierarchické shlukování

Pokud chceme rozdělit objekty do k shluků, můžeme k tomu použít některou z hierarchických metod a vybrat si tu úroveň výsledného binárního stromu, která má právě k shluků. Nevýhodou takového postupu je skutečnost, že u hierarchického shlukování nemůžeme změnit výsledek předchozího kroku. Proto v takovýchto situacích upřednostňujeme použití nehierarchických metod, které dávají lepší výsledky. Jedním z nejjednodušších, nejrychlejších a i nejpoužívanějších algoritmů je metoda k-průměrů (k-means).

5.2.1 Metoda k-průměrů

Máme daný požadovaný počet shluků k. Nejprve si zvolíme k reprezentantů $(t_1, t_2, \ldots, t_k)$ pro shluky, které budeme vytvářet. Tito reprezentanti mohou, ale nemusí, pocházet ze zadaných objektů a můžeme je tedy volit libovolně podle našeho úsudku, viz [3]. Pokud nemáme žádnou domněnku, jak by těchto k bodů mělo vypadat, je možné je získat náhodným výběrem ze vstupních dat. Reprezentanty nazýváme centroidy a iteračně je upravujeme tak, abychom minimalizovali výraz

$$\sum_{j=1}^{k} \sum_{\boldsymbol{X}_i \in C_j} d\left(\boldsymbol{X}_i, t_j\right). \tag{5.1}$$

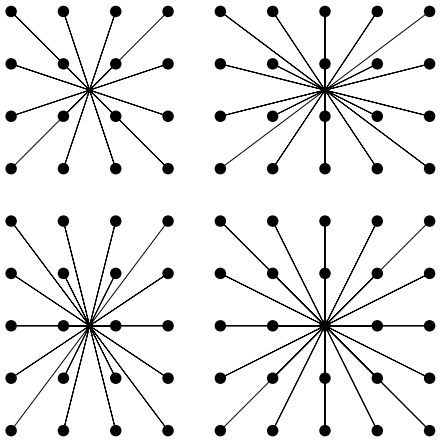
Každá iterace se skládá ze dvou částí. V první části třídíme objekty do shluků. Objekt $\boldsymbol{X}_i$ přiřadíme do shluku Cj, kde $j \in \{1, 2, \ldots, k\}$ je takový index, pro který výraz $d(\boldsymbol{X}_i, t_j)$ nabývá svého minima. V druhé části vypočítáme nové centroidy

$$t_i = \min_{t_i} \sum_{\boldsymbol{X}_i \in C_j} d\left(\boldsymbol{X}_i, t_j\right).$$

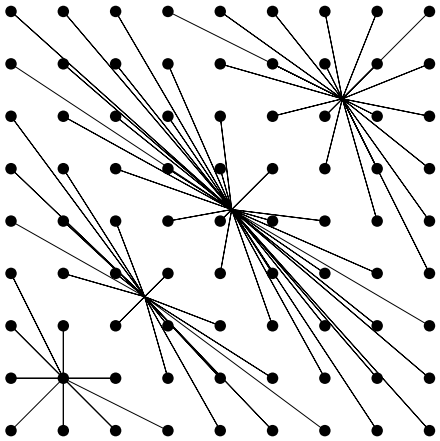
Algoritmus vede k minimalizaci výrazu (5.1) podle proměnných $\boldsymbol{X}_i$ v první části a podle proměnných t_j v části druhé. Pokud po poslední iteraci získáme stejné shluky jako v po iteraci předchozí, algoritmus končí.

Po ukončení algoritmu získáme rozdělení do shluků takové, které je lokálním minimem výrazu (5.1). Různě zvolené počáteční body nás tedy mohou dovést k jiným výsledkům. Tato metoda je citlivá na odlehlá pozorování, protože ta mohou výrazně poznamenat napočítávané centroidy. Nevýhodou metody je nutnost definovat výsledný počet shluků k ještě před samotným začátkem algoritmu.

Obrázky 5.1 a 5.2 vytvořené pomocí systému Mathematica (dle předlohy v [8]) ukazují, jak počáteční volba reprezentativních bodů může ovlivnit výsledek. Oba výsledné rozklady leží v lokálním minimu kriteriální funkce metody k-průměrů.



Obrázek 5.1



Obrázek 5.2

5.2.2 Příklad

Použijeme metodu k-průměrů, abychom si ověřili, zda při $k = 2$ dostaneme stejný výsledek jako u hierarchické metody nejbližšího souseda. Zvolíme si počáteční body (inspirované výsledkem hierarchické metody) například takto:

$$t_1 = \boldsymbol{X}_1 = 2,$$
$$t_2 = \boldsymbol{X}_5 = 5, 2.$$

Nyní na základě euklidovské vzdálenosti od výše definovaných centroidů zařadíme objekty do shluků.

$$C_1 = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_4\}$$
$$C_2 = \{\boldsymbol{X}_3, \boldsymbol{X}_5, \boldsymbol{X}_6\}$$

Nově si přepočítáme centroidy

$$t_1 = \frac{\boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_4}{3} = 2, 6,$$
$$t_2 = \frac{\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6}{3} = \frac{19, 7}{3}.$$

Znovu zařadíme všechny objekty do dvou skupin podle nových centroidů.

$$C_1 = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_4\}$$
$$C_2 = \{\boldsymbol{X}_3, \boldsymbol{X}_5, \boldsymbol{X}_6\}.$$

Získali jsme stejný rozklad jako v předchozím kroku, algoritmus tedy končí. Výsledek metody k-průměrů se v tomto příkladě shoduje s výsledkem metody nejbližšího souseda.

5.2.3 Odhad vhodného počtu shluků

Pokud nemáme o datech takovou informaci, ze které bychom byli schopni určit vhodný počet shluků k, nezbývá nám než k odhadnout. K tomu můžeme použít hrubá doporučení, například $k = \sqrt{\frac{n}{2}}$, nebo si vytvořit několik rozkladů s různým počtem shluků a z nich vybrat ten vhodný (ovšem jaký počet je "vhodný", nemusí být z výsledků jednoznačně patrné). Existuje více exaktních metod pro určení vhodného k, žádná z nich však není jednoznačně nejlepší.

Jednou z metod, která v praxi dosahuje dobrých výsledků, je metoda založená na *silhouette width* (viz [6]). Hodnotu silhouette width spočítáme jako:

$$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{\max\left\{a_{(i)}, b_{(i)}\right\}},\ i \in \{1, 2, 3, \ldots, n\},$$

kde $a_{(i)}$ je průměrná míra nepodobnosti i-tého objektu ke všem ostatním objektům ve stejném shluku a $b_{(i)}$ je minimum ze všech průměrných vzdáleností mezi i-tým objektem a všemi objekty jiného pevného shluku (tedy průměrná vzdálenost $\boldsymbol{X}_i$ s prvky "nejbližšího" shluku). $s_{(i)}$ nabývá hodnot z intervalu $\langle -1, 1\rangle$, kde hodnoty blízké 1 znamenají, že objekt je přidělen správnému shluku, objekt s hodnotami kolem 0 se patrně nachází v "prostoru mezi shluky" a u hodnot blízkých -1 jde pravděpodobně o zařazení do špatného shluku.

Za vhodné k považujeme takové číslo, pro které průměrná hodnota silhouette width nabývá svého maxima

$$\max \bar{s} = \max \sum_{i=1}^{n} \frac{s_{(i)}}{n}.$$

Kapitola 6

Shluková analýza společností pražské burzy

Shluková analýza nám může dát odpověď na otázku, zda společnosti, které považujeme za podobné, protože podnikají ve stejném odvětví, jsou si podobné i ve vybraných finančních ukazatelích. Naopak u společností, které považujeme za značně rozdílné, očekáváme, že tato rozdílnost bude patrná i ve výsledku shlukové analýzy.

Zásadní význam má samozřejmě už samotná volba finančních ukazatelů. Snažíme se vybrat jen takové, které jsou pro nás relevantní. Pokud se výsledek shlukové analýzy nebude shodovat s naším intuitivním roztříděním společností, můžeme pátrat po příčinách. Těmi mohou být například nevhodná volba vstupních dat nebo případná neexistence podobností firem v takové formě, jakou jsme uvažovali.

Jako vstupní data použijeme veřejně dostupné údaje nebo jejich odhady známé ke dni 23.7.2009 (viz tabulku 6.1. Zdroj: [9], internetové stránky jednotlivých společností). U každé firmy sledujeme 5 ukazatelů: zadluženost (poměr dluh/aktiva), velikost (tržní kapitalizace v USD), výše dividendy na akcii dělenou cenou akcie, ziskovost (zisk na akcii dělený cenou akcie) a poměr účetní a tržní hodnoty.

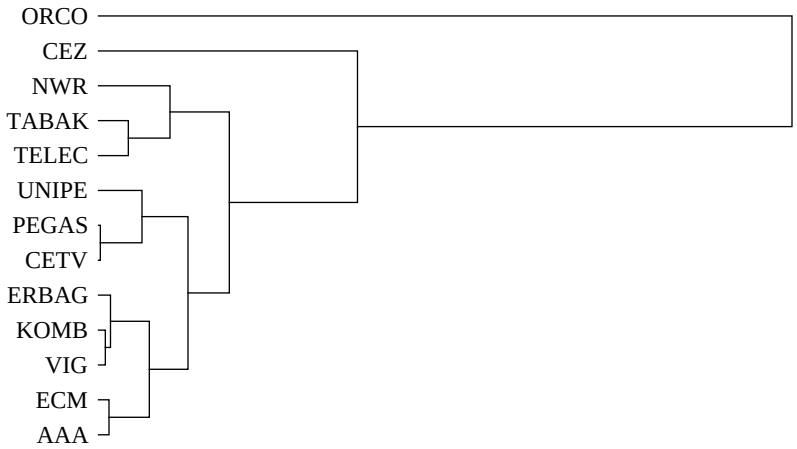
	dluh	tržní kapitalizace	dividenda	EPS/cena	úč./tržní cena
AAA	0,90	33 132 705	0,0000	-0, 0383	0,0164
CETV	0,58	1 253 117 198	0,0000	-0, 3613	0,8012
CEZ	0,61	26 422 525 664	0,0563	0, 1061	0,3627
ECM	0,86	117 241 659	0,0000	-0, 0759	0,0353
ERBAG	0,89	9 510 460 301	0,0316	0, 1324	1,2367
KOMB	0,91	5 986 133 079	0,0654	0, 1247	0,5915
NWR	0,71	1 380 674 167	0,1239	0, 2370	0,6598
ORCO	0,83	99 690 422	0,2263	-6, 4522	4,4997
PEGAS	0,60	190 076 850	0,0000	0, 0766	0,7482
TABAK	0,38	738 712 719	0,1277	0, 0894	0,6028
TELEC	0,25	8 243 798 506	0,1093	0, 0804	0,5304
UNIPE	0,34	1 128 349 578	0,0000	0, 0031	1,8141
VIG	0,88	5 677 123 968	0,0359	0, 1800	0,9884

Tabulka 6.1

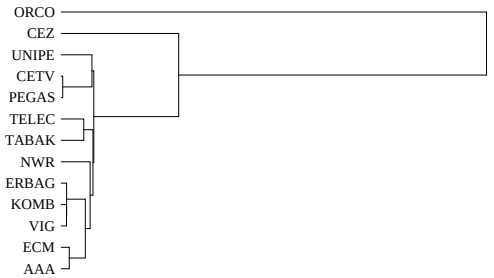
Protože data jsou vyjádřena ve velmi odlišných jednotkách, je vhodné je před samotným začátkem shlukování standardizovat. Sloupce s jednotlivými proměnými jsou pro následující výpočty upraveny tak, aby měly nulovou střední hodnotu a jednotkový rozptyl.

Musíme ještě rozhodnout, jakou metodu použijeme. V tomto případě budeme vybírat z hierarchických metod, i když použití některé z nehierarchických metod by samozřejmě možné bylo. Vzhledem k malému množství firem by bylo obtížné vhodně určit počet shluků, protože společnosti obchodované na pražské burze nelze dost dobře do dostatečně velkých homogenních shluků rozdělit. Hierarchické metody nám navíc umožní výsledek přehledně demonstrovat v podobě dendrogramu.

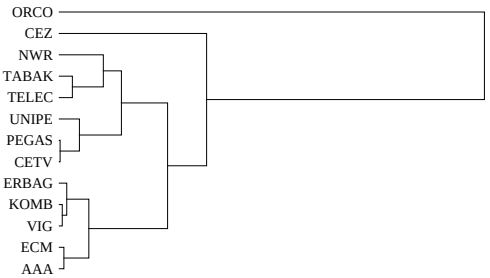
Výsledek závisí na volbě míry nepodobnosti a na použité metodě. Mírou nepodobnosti je v tomto výpočtu euklidovská vzdálenost a vzdálenost mezi shluky je určována podle principu průměrné vzdálenosti prvků (obrázek 6.1 zachycuje výsledný dendrogram vytvořený systémem Mathematica). Pro srovnání jsou zde znázorněné i dendrogramy vytvořené pomocí metody nejbližšího respektive nejvzdálenějšího souseda (obrázky 6.2 a 6.3).



Obrázek 6.1



Obrázek 6.2: metoda nejbližšího souseda



Obrázek 6.3: metoda nejvzdálenějšího souseda

Ačkoliv výše zobrazené dendrogramy nejsou úplně totožné, není v nich žádný zásadní rozdíl. Na první pohled je výrazný odstup společnosti Orco od všech ostatních společností kótovaných na pražské burze. Především absence podobnosti s druhým developerem ECM je překvapující.

ČEZ je od ostatních firem vzdálen zejména díky své velikosti (tuto domněnku potvrdí i výpočet s totožnými daty bez údajů o tržní kapitalizaci). Následuje několik skupin obsahujících dvě až tři společnosti, které se svými finančními ukazateli očividně podobají. Zatímco podobnost mezi společnostmi Telefónica O2 a Philip Morris ČR může i přes různou oblast podnikání spočívat ve skutečnosti, že jde o stabilní společnosti standardně vyplácející relativně vysoké dividendy a označované jako defenzivní akcie, podobnost firem Pegas Nonwovens a Central European Media Enterprises nebo již zmiňovaného developera ECM a AAA takto jednoduše zdůvodnit nelze. Na druhou stranu trojice finančních skupin Erste, VIG a Komerční banky opravdu tvoří relativně homogenní skupinu.

Kapitola 7

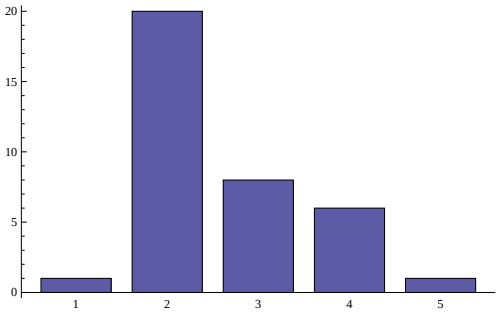
Shluková analýza vybraných společností newyorské burzy

V této kapitole budeme pracovat s některými ze společností kótovaných na NYSE a jejich finančními ukazateli dostupnými v systému Wolfram Mathematica. Z firem podnikajících v pěti různých odvětvích (bankovnictví, ropný průmysl, telekomunikace, farmacie a software) vybereme ty, ke kterým má Mathematica informace o třech vybraných ukazatelích. Těmito ukazateli jsou roční dividenda dělená současnou cenou akcie, poměr ceny akcie a zisku na akcii (P/E) a poměr tržní a účetní hodnoty. Protože k některým společnostem nemá Mathematica všechny tyto informace k dispozici, je počet použitých společností nižší, než počet všech společností z výše vybraných odvětvích obchodovaných na NYSE. Každá z pěti skupin určených různými odvětvími má však stále uspokojivý počet reprezentantů (7–36).

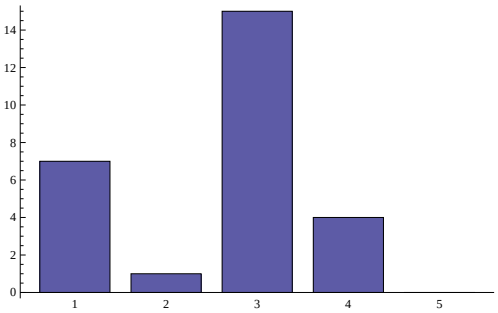
Budeme se snažit nalézt odpověď na otázku, zda skupiny firem tvořené odvětvími se shodují se skupinami vzniklými na základě podobnosti třech vybraných finančních ukazatelů. Na rozdíl od předchozí kapitoly použijeme nyní nehierarchickou metodu. Známe totiž přesný počet odvětví a tedy i počet shluků, které chceme vytvořit.

Stejně jako v předchozích příkladech data standardizujeme, abychom zabránili přílišnému převážení některého z ukazatelů. Ke všem následujícím výpočtům je opět použit systém Mathematica se zachováním výchozího nastavení - míra nepodobnosti je určena euklidovskou vzdáleností a algoritmus je založený na metodě k-průměrů. Vzhledem k vyššímu rozsahu zde nejsou uvedena vstupní data této úlohy, zájemce je však najde na přiloženém CD, a to včetně všech výpočtů.

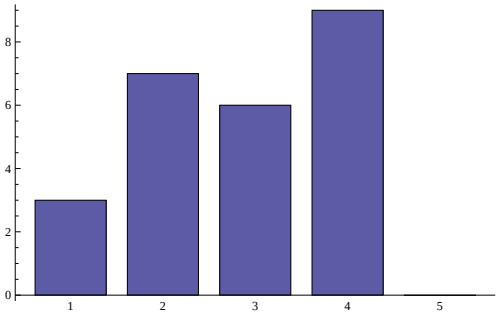
Na obrázcích 7.1 až 7.5 je zachyceno, kolik firem z jednotlivých odvětví bylo zařazeno do shluků 1, 2, 3, 4 a 5. Z prvního obrázku tedy plyne, že jedna banka byla zařazena do shluku 1, 20 bankovních společností bylo přiřazeno do shluku 2 atd.



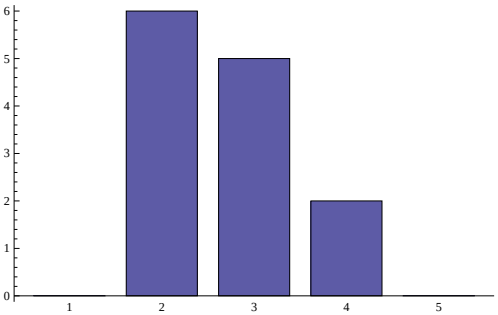
Obrázek 7.1: bankovní společnosti



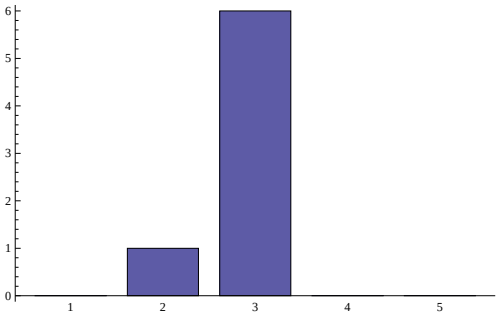
Obrázek 7.2: ropné společnosti



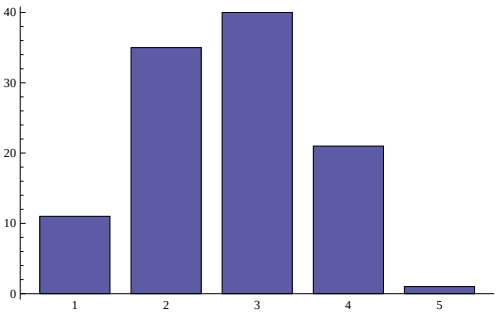
Obrázek 7.3: telekomunikační společnosti



Obrázek 7.4: farmaceutické společnosti



Obrázek 7.5: softwarové společnosti



Obrázek 7.6: velikosti shluků

Přestože v případě bankovních, ropných nebo softwarových společností lze najít shluk, do kterého nadpoloviční většina firem spadá, souvislost mezi původním rozdělením podle odvětví a výslednými shluky není valná. Tím spíše, že u ropných i softwarových společností je převládající shluk shodný.

Z obrázků je patrné, že poslední shluk je jednoprvkový. Tvoří ho jediná společnost – jakési odlehlé pozorování. I pokud takovéto výrazné společnosti ze vstupních dat odstraníme nebo pokud zvýšíme počet shluků, abychom z odlehlých pozorování vytvořili jednoprvkové shluky a takto se jich "zbavili", provázanost mezi odvětvími a výslednými shluky se nezlepší.

Jedním z možných vysvětlení relativně nízkého provázání mezi odvětvími a vytvořenými shluky je fakt, že výše použité proměnné zvolené společnosti nemusejí dostatečně dobře popisovat. Mathematica nedisponuje všemi údaji o společnostech

obchodovaných na NYSE a proto některé údaje (například o zadlužení) nebyly použity.

Druhým možným vysvětlením je, že firmy zařazené do určitého odvětví opravdu tvoří nehomogenní skupinu. Ostatně takovýto závěr se objevil už v článku [1], ve kterém ovšem byly společnosti porovnávány na základě korelace finančních ukazatelů během desetileté periody a ne podle jejich aktuální hodnoty. Pokud je hypotéza nehomogenity správná, tak například diverzifikace portfolia vzhledem k sektorům nemusí znamenat dostatečné rozložení rizika, pokud chybí odpovídající diverzifikace uvnitř sektorů. Tímto tedy shluková analýza podpírá myšlenku, která je v mnoha publikacích (např. v [5]) zdůvodněná obvykle jen historickým vývojem.

Literatura

[1] Dahlstedt, R. a kol.: *On the usefulness of standard industrial classifications in comparative financial statement analysis*. European Journal of Operational Research 79 (1994). 230-238.

[2] Dupačová, J., Hurt, J., Štěpán, J.: *Stochastic Modeling in Economics and Finance*. Kluwer Academic Publishers. Dordrecht 2002.

[3] Hurt, J.: *Mnohorozměrná statistická analýza*. Přednáška na MFF UK, ZS 2008/2009.

[4] Kelbel, J., Šilhán, J.: *Shluková analýza*. Praha, 2002. Dostupný na http://gerstner.felk.cvut.cz/biolab/X33BMI/slides/KMeans.pdf.

[5] Kohout, P.: *Investiční strategie pro třetí tisíciletí*. Grada 2005.

[6] Rao, A. Ramachandra, Srinivas, V.V.: *Regionalization of Watersheds: An Approach Based on Cluster Analysis*, Dordrecht 2008.

[7] Řezanková, H.: *Klasifikace pomocí shlukové analýzy*. Pardubice, 2004. Dostupný na http://nb.vse.cz/~REZANKA/Shlukova_analyza2003.pdf.

[8] Šarmanová, J.: *Metody dolování znalostí z dat*. In Datakon 2002. Ed. Chlápek, D. Ostrava, 2002.

[9] webové stránky společnosti Patria Online, a.s., *http://www.patria.cz/*.

[10] Wolfram, S.: Mathematica v. 6.0.3. Help/tutorial/PartitioningDataIntoClusters.

[11] Žák, L.: *Shluková analýza I.* Časopis Automatizace. Ročník 47, číslo 3, březen 2004.