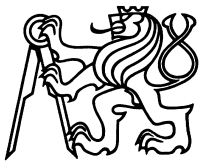
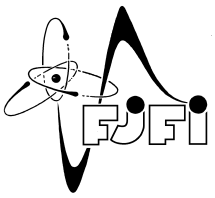


České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



Aplikace klasifikačních metod na egyptologická data

Application of classification methods on egyptoligical data

Bakalářská práce

Autor: **Jazmína Kreanová**

Vedúci práce: **Ing. Marek Bukáček, Ph.D.**

Akademický rok: 2021/2022

České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská

Katedra: matematiky

Akademický rok: 2021/2022

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student:	Jazmína Kreanová
Studijní program:	Aplikace přírodních věd
Studijní obor:	Matematické inženýrství
Studijní zaměření:	Aplikované matematicko-stochastické metody
Název práce (česky):	Aplikace kasifikačních metod na egyptologická data
Název práce (anglicky):	Application of classification methods on egyptological data

Pokyny pro vypracování:

1) Proveďte rešerši základních statistických metod s důrazem na klasifikační nástroje.
2) Implementujte vybrané varianty k-mean algoritmu a ilustrujte jejich vlastnosti na testovacích datech.
3) Seznamte se s daty poskytnutými Českým egyptologickým ústavem Filozofické fakulty Univerzity Karlovy.
4) S použitím implementovaného algoritmu pro klasifikaci ve vícerozměrném diskrétním prostoru analyzujte tituly staroegyptské společnosti.

Doporučená literatura:

1) L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis. John Wiley \& Sons, USA, 2009.
2) J. Anděl, Základy matematické statistiky, Matfyzpress, Praha, 2011.
3) D. Jones, An index of ancient Egyptian titles, epithets and phrases of the Old Kingdom. Archaeopress, Oxford, 2000.

Jméno a pracoviště vedoucího bakalářské práce:

Ing. Marek Bukáček
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, České vysoké učení technické v Praze, Trojanova 13, 120 00 Praha 2

Jméno a pracoviště konzultanta:

Datum zadání bakalářské práce: 31.10.2021

Datum odevzdání bakalářské práce: 7.7.2022

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 21. října 2021

.....
garant oboru

.....
vedoucí katedry

.....
děkan

Poďakovanie:
Na tomto mieste by som sa chcela poďakovať svojmu školiteľovi Ing. Markovi Bukáčkovi, Ph.D. za jeho nekonečnú trpezlivosť, odbornosť, ochotu, srdečný prístup a za každú cennú radu, ktorou smeroval vývoj mojej bakalárskej práce. Osobitné poďakovanie patrí mojej rodine a blízkym za to, že mi boli neustálou oporou počas štúdia a pomáhali mi zdolávať prekážky.

Čestné prehlásenie:
Prehlasujem, že som túto prácu vypracovala samostatne a uviedla som všetku použitú literatúru.

V Prahe dňa 5. januára 2022 Jazmína Kreanová

Názov práce:

Aplikace klasifikačních metod na egyptologická data

Autor: Jazmína Kreanová

Obor: Matematické inženýrství

Zameranie: Aplikované matematicko-stochastické metody

Druh práce: Bakalářská práce

Vedúci práce: Ing. Marek Bukáček,Ph.D. České vysoké učení technické v Praze, Fakulta jaderná a fyzikálne inženýrská, Katedra matematiky

Abstrakt: Táto práca sa zaoberá zhlukovou analýzou za účelom jej aplikácie na egyptologické dáta z archeologického výskumu staroegyptskej spoločnosti. Uvádza základné princípy zhlukovania, delenie metód a popisuje vybrané algoritmy. Hlavná časť sa venuje predovšetkým nehierarchickým metódam, konkrétne metóde k-means a jej podobným. Uvádzajú sa dva nástroje na určenie počtu zhlukov metóda kolena a obrysový koeficient. Získané znalosti sa využívajú pri demonštrácií na simulovanom dvojrozmernom datasete. Na záver sa použijú algoritmy k-means a k-modes na experimentálne spracovanie dát získaných z priečelí hrobiek Starej ríše Egypta, ktoré poskytol Český egypologický ústav.

Kľúčové slová: dvojfázový k-means, k-means, k-means++, k-medoids, k-modes, metóda kolena, miera podobnosti, obrysový koeficient, staroegyptská spoločnosť, zhluk, zhluková analýza

Title:

Application of classification methods on egyptological data

Author: Jazmína Kreanová

Abstract: This thesis deals with cluster analysis for a purpose of its application on egyptological data of archaeological research of the Ancient Egypt society. The principles of clustering and types of methods are presented and described. The main part deals with partitional clustering, especially with k-means and other similar algorithms. Two ways of determining the number of clusters are given, these being the elbow rule and the silhouette coefficient. The conducted research is demonstrated on a simulated two-dimensional data. The final part uses k-means and k-modes algorithms to analyse data of the Old Kingdom of Egypt, which are provided by the Czech Institute of Egyptology.

Key words: ancient Egypt society, cluster, cluster analysis, elbow rule, k-means, k-means++, k-medoids, k-modes, silhouette coefficient, similarity measure, two-phase k-means

Obsah

Úvod

Ľudia za svoj život zanechávajú vo svete stopy bez toho, aby vedeli či o tisícky rokov niekto bude skúmať kým boli a ako ovplyvnili svoju dobu. Český egyptologický ústav Univerzity Karlovej, za desiatky rokov vykopávok a výskumov, nazhromaždil rozsiahle množstvo dát, ktoré poskytujú šancu nahliadnuť do histórie Starej ríše Egypta, odkrývať stopy, ktoré za sebou ľudia zanechali, skúmať štruktúru spoločnosti, v ktorej žili.

Táto bakalárska práca sa venuje matematickému aspektu skúmania starovekého Egypta v spolupráci s českými egyptológmi. Z množstva nástrojov, ktoré nám matematika a strojové učenie ponúkajú, sa v tejto práci zaoberáme zhlukovou analýzou (z angl. *Cluster analysis*), aby sme mohli vďaka získaným poznatkom z priečelí hrobiek hľadať a skúmať skupiny či trendy, ktoré sa počas etáp ríše formovali. Tento smer nám ďalej umožňuje rozčleňovať a klasifikovať spoločnosť a tým sa priblížiť k odhaleniu historických súvislostí, zmien spoločenských pomerov a fungovania spoločnosti.

V prvej kapitole tejto práce predstavíme pojem a základy zhlukovej analýzy, jej obecné princípy a postupy. Ukážeme delenie zhlukovacích metód a ich pravidlá. Pri používaní zhlukovacích nástrojov je dôležité určiť miery podobností dát, vymenujeme preto niekoľko možností. Nakoniec popíšeme základ vybraných zhlukovacích prístupov a spôsob ich fungovania.

Prostredníctvom druhej kapitoly sa zoznámime s algoritmom k-means a jeho použitím. Uvádzame všeobecný popis algoritmu. Ďalej popíšeme niektoré algoritmy, ktoré sú založené na báze algoritmu k-means, pričom sa adaptujú v rôznych smeroch efektivity, konkrétne prezentujeme: k-means++, k-medoids, dvojfázový k-means a k-modes.

Tretia kapitola je zameraná na aplikáciu popísaných metód na umelo vytvorené dvojrozmerné dáta. Priblíži správanie algoritmov a podobu výsledkov. Ukážeme dva spôsoby určenia vhodného počtu zhlukov a vlastnosti algoritmov spolu s vplyvom voľby mier podobnosti.

V poslednej časti sa zoznámime s dátami poskytnutými českým egyptologickým ústavom a následne aplikujeme vhodné nástroje zhlukovej analýzy.

Kapitola 1

Úvod do zhlukovej analýzy

Zhluková analýza (z ang. *Cluster analysis*) je jednou zo štatistických metód (resp. metód *machine learningu*), ktorá sa zaoberá zoskupovaním sebe podobných dát. Prvýkrát pojem zhlukovej analýzy sformuloval profesor psychológie R.C. Tyron v roku 1939. Zhlukovú analýzu interpretoval ako všeobecný logický postup formulovaný ako procedúra, pomocou ktorej objektívne zoskupujeme jedincov do skupín na základe ich podobnosti a rozdielnosti. V tejto práci budeme okrem Tyronovej logiky vychádzať predovšetkým z [2], [3], [4] a [10].

Cieľom zhlukovej analýzy je zaradiť prvky respektíve objekty do zhlukov resp. tried tak, aby si prvky nachádzajúce sa v rovnakom zhluku boli viac podobné než prvky z rozličných zhlukov. Každý objekt je popísaný pomocou charakteristických čŕt, znakov, ktoré umožňujú určiť podobnosť medzi dvomi objektmi. Napríklad ak naším objektom je konkrétny Staroegypťan, medzi charakteristické znaky, ktoré ho popisujú patria napríklad jeho titul, pohlavie, obdobie života a podobne. Na základe týchto znakov môžeme sledované dáta zaradiť do tried, ktoré umožňujú jednoduchší popis. Pri aplikácii metód zhlukovej analýzy sa zameriavame na vybrané znaky, ktoré vystupujú ako premenné. Môžu byť rôzneho typu: číselné, ordinálne, slovné, binárne apod. Typicky sledujeme viacero premenných, tým pádom ide o viacdimenzionálnu charakterizáciu tried.

Základným predpokladom je jednoznačné zaradenie objektu, tj. objekt patrí práve do jedného zhluku. Hovoríme o *disjunktnom zhlukovaní*. V druhom prípade, ak objekt patrí viac ako jednému zhluku, ide o *prekrývajúce sa zhlukovanie*.
V oblasti zhlukovej analýzy sa stretávame s množstvom algoritmov, ktorých základom môžu byť rôzne princípy. Rozlišujeme dva základné prístupy:

- učenie s učiteľom (*supervised learning*) a
- učenie bez učiteľa (*unsupervised learning*).

Typickým faktorom učenia s učiteľom je mať k dispozícii informáciu o zaradení objektov k vopred známym skupinám. Úlohou je vytvoriť (naučiť) model tak, aby na jeho základe bolo možné zaradiť aj také objekty, o ktorých príslušnosť do zhlukov nie je na začiatku procesu známa. Naopak, pri učení bez učiteľa zaradenie objektov či počet zhlukov nie sú známe faktory. Cieľom je v tomto prípade klasifikovať všetky prvky a začleniť ich do príslušných zhlukov vhodného charakteru a počtu.

V tejto práci sa budeme zaoberať učením bez učiteľa. Pod pojmom *zhlukovanie* budeme rozumieť postup, kedy vopred nie je známy počet skupín a ani príslušnosť akéhokoľvek objektu. Cieľom tohto postupu bude klasifikovať všetky objekty analýzy a výsledok vhodne interpretovať.

Majme na pamäti, že rôzne algoritmy môžu viesť k nájdeniu iného vzoru v dátach, pretože typy objektov sa utvárajú až v procese zhlukovania.

Postup zhlukovania môžeme popísať v štyroch dôležitých krokoch:

1. Voľba miery podobnosti dát
2. Voľba metódy zhlukovej analýzy
3. Určenie počtu zhlukov
4. Interpretácia výsledkov

V štatistickej analýze dát (alebo machine learningu obecne) je vhodné vstupné dáta používať vo forme dátovej matice s rozmermi $n \times m$, označme ju $\mathbf{X}$ s prvkami x_{ij}, kde $i = 1, \ldots, n$ a $j = 1, \ldots, m$. Základ teda tvoria m-rozmerné dáta s počtom n.
K dispozícií máme n vektorov v riadkoch matice. Každý vektor reprezentuje objekt. Jednotlivé stĺpce matice, resp. zložky vektora, nesú informácie o konkrétnom štatistickom znaku. Tvoria premenné. Pre $i \in \widehat{n}$ platí $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, kde x_{ij} predstavuje hodnotu j-teho znaku i-teho objektu a $j \in \widehat{m}$.
Ďalším typom vstupných dát je kontingenčná tabuľka (dvojrozmerná tabuľka združených početností) pre dve nominálne premenné. Táto práca sa však vstupnými dátami vo forme kontingenčných tabuliek nevenuje.

1.1 Miery podobnosti

Prvým pilierom zhlukovej analýzy je určenie podobnosti dvoch objektov. Ako sme už zmienili, našou úlohou je utvoriť zhluky objektov, ktoré sú si najviac podobné. Aby sme mohli túto podobnosť určiť, musíme správne zvoliť spôsob, akým budeme porovnávať objekty medzi sebou. K dispozícií máme miery podobnosti a miery vzdialenosti. Výber závisí na type znakov popisujúcich objekty.

Ak máme k dispozícií kvantitatívne dáta, môžeme vzťah medzi dvomi objektmi popísať mierami vzdialenosti. Dáta teda vnímame ako priestorové útvary.

Definícia 1.1.1. Majme zobrazenie $\rho : \mathbb{R}^m \to \mathbb{R}$ pre ktoré platí:

1. $\rho(x, y) \geq 0 \quad \text{a} \quad \rho(x, y) = 0 \quad \Leftrightarrow \quad x = y \quad \text{pre} \quad \forall x, y \in \mathbb{R}^m$,
2. $\rho(x, y) = \rho(y, x) \quad \forall x, y \in \mathbb{R}^m$,
3. $\rho(x, y) + \rho(y, z) \geq \rho(x, z) \quad \forall x, y, z \in \mathbb{R}^m$.

Potom o zobrazení ρ povieme, že je *metrikou* [12].

V tejto práci budeme označovať vzdialenosť dvoch prvkov ako $d(x, y)$ podľa anglického *distance* a v zhode s použitou literatúrou. Ďalej za *mieru* označíme také zobrazenie $\rho : \mathbb{R}^m \to \mathbb{R}$, ktoré spĺňa prvé dva body definície 1.1.1.

Spomenieme niektoré z najčastejšie používaných mier vzdialeností [4]:

Minkowského vzdialenosť

$$d_{Mi}(x_i, x_k) = \sqrt[q]{\sum_{j=1}^{m}(x_{ij} - x_{kj})^q}, \tag{1.1}$$

bloková (Manhattanská) vzdialenosť, ktorá je špeciálnym príkladom Minkowského vzdialenosti pre $q = 1$

$$d_B(x_i, x_k) = \sum_{j=1}^{m} |x_{ij} - x_{kj}|, \tag{1.2}$$

euklidovská vzdialenosť, ktorá je Minkowského konkrétnym prípadom pre $q = 2$

$$d_E(x_i, x_k) = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2}, \tag{1.3}$$

logaritmická vzdialenosť

$$d_L(x_i, x_k) = log(1 + \sum_{j=1}^{m}(x_{ij} - x_{kj})^2), \tag{1.4}$$

Mahalanobisova vzdialenosť[1]

$$d_M(x_i, x_k) = \sqrt{(x_i - x_k)^T \mathbb{C}^{-1}(x_i - x_k)}, \tag{1.5}$$

kde $\mathbb{C}$ je kovariančná matica, ktorá do výpočtu zahŕňa aj väzbu medzi objektmi. Využívame ju ak body v priestore nie sú ortogonálne.

Ďalšou možnosťou popisu vzťahu dvoch objektov je miera podobnosti resp. miera nepodobnosti.
Miera podobnosti $p \in \langle 0; 1\rangle$ popisuje hodnotou 0 prvky, ktoré sú maximálne rozdielne a hodnotou 1 prvky totožné. Inak tomu je u mier vzdialenosti, ktoré popisujú podobnostný vzťah opačným spôsobom, 0 pre totožné prvky a s rastúcou vzdialenosťou rastie odlišnosť. Z toho dôvodu sa niekedy používa *miera nepodobnosti*.

Lemma 1.1.1. Každú mieru podobnosti možno previesť na mieru nepodobnosti.

Dôkaz: Ak $p \in \langle 0; 1\rangle$ je ľubovoľná miera podobnosti, potom mieru nepodobnosti možno vytvoriť komplementárnou operáciou vlastnou každej ohraničenej podobnosti.
Mieru nepodobnosti teda možno dodefinovať nasledovne:

$$\rho_N = 1 - p. \tag{1.6}$$

□

Miera nepodobnosti teda zastupuje rovnaký princíp ako miery vzdialenosti. V literatúre nájdeme značenie S resp. D pre mieru podobnosti resp. nepodobnosti. Budeme značiť ako D.

Medzi využívané miery podobnosti patrí napríklad Kosínova miera, Jaccardov koeficient, Diceho koeficient či Czekankowského koeficient [4].
Ak máme k dispozícii vlastnosti rôznych dátových typov, je vhodné použiť na to uspôsobenú mieru. Tou je napríklad **Gowerov koeficient podobnosti** [4]

$$S_{ij} = \frac{\sum_{l=1}^{m} w_{ijl} S_{ijl}}{\sum_{l=1}^{m} w_{ijl}}, \tag{1.7}$$

kde S_{ijl} je miera podobnosti medzi objektmi x_i a x_j na základe l-tej premennej (l-tého znaku) a w_{ijl} popisuje prítomnosť znaku v premenných. Ak x_{il} alebo x_{jl} chýba, alebo je nulová, potom $w_{ijl} = 0$. Ak je tomu inak $w_{ijl} = 1$.

1.2 Podobnosti premenných

Ak posudzujeme vzťah dvoch premenných (čŕt), hovoríme o *závislosti*. Miery podobnosti označujú miery štatistickej závislosti. Výberový **Pearsonov korelačný koeficient** [13] je braný ako základná miera pre posudzovanie závislosti dvoch kvantitatívnych premenných. Pre k-tu a l-tú premennú ho spočítame podľa vzorca

$$r_{kl} = \frac{s_{kl}}{s_k s_l} = \frac{\sum_{i=1}^{n}(x_{ik} - \overline{x_k})(x_{il} - \overline{x_l})}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x_k})^2 \sum_{i=1}^{n}(x_{il} - \overline{x_l})^2}}, \tag{1.8}$$

kde $\overline{x_k}$ resp. $\overline{x_l}$ je aritmetický priemer hodnôt k-tej resp. l-tej premennej. Pre korelačný koeficient platí $r_{kl} \in \langle -1; 1 \rangle$, pričom pre $r_{kl} = 0$ nastáva lineárna nezávislosť. Hodnoty -1 a 1 ekvivalentne ukazujú maximálny súhlas medzi dvomi premennými a silná korelácia ná umožňuje brať do úvahy len jednu z týchto premenných a redukovať príznakový priestor. Mieru nepodobnosti premenných teda spočítame vo forme $D_{kl} = 1 - r_{kl}^2$ alebo $D_{kl} = 1 - |r_{kl}|$.

1.3 Klasifikácia metód zhlukovej analýzy

Metódy zhlukovej analýzy bývajú klasifikované na základe rôznych pohľadov, podľa toho k akému cieľu metódy aspirujú. Najzákladnejším typom klasifikácie je klasifikácia podľa spôsobu formovania zhlukov:

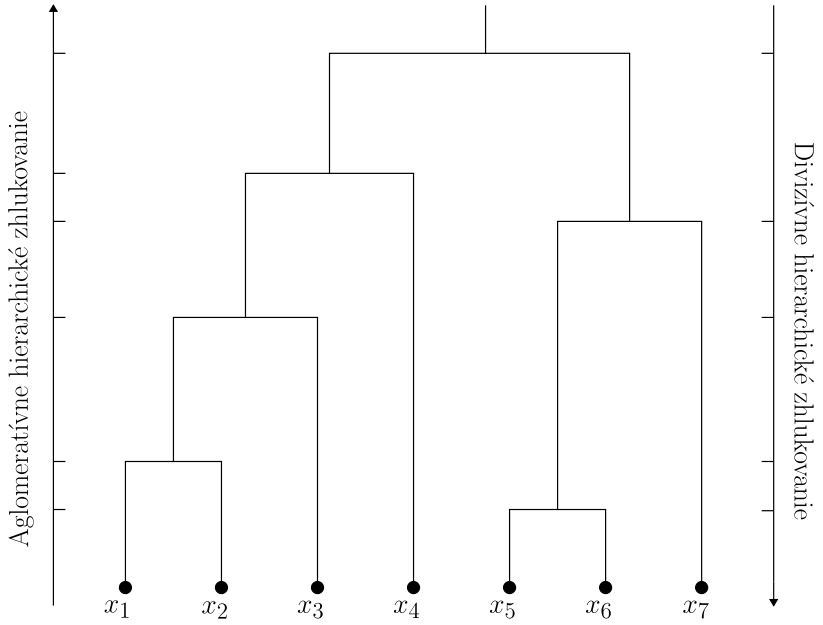
- hierarchické metódy,
- nehierarchické metódy (metódy rozkladu).

1.3.1 Hierarchické metódy

Pre hierarchické zhlukovanie je charakteristická sekvenčná (hierarchická) štruktúra procesu zhlukovania. Na rozdiel od nehierarchického zhlukovania nie je potrebné vopred poznať konkrétny počet zhlukov či uvažovať špecifickú počiatočnú inicializáciu procesu (rozumieme tým voľbu počiatočných pseudocentier, okolo ktorých sa formujú a optimalizujú zhluky). V hierarchickom procese zhlukovania uvažujeme dva smery: zdola-nahor, zhora-nadol. Buď začíname

stavom, kedy každý objekt predstavuje samostatný zhluk a cieľom je postupným zlučovaním najpodobnejších zhlukov dosiahnuť, že všetky objekty sa nachádzajú v jednom spoločnom zhluku. Tento spôsob sa nazýva *aglomeratívne zhlukovanie*. V opačnom smere teda začíname stavom, kedy všetky objekty náležia spoločnému zhluku a postupným rozkladom sformujeme n zhlukov po jednom prvku. Hovoríme o *divizívnom zhlukovaní*. V procese pracujeme s maticou vzdialeností respektíve nepodobností a na jej základe činíme rozhodnutia o spojení či rozklade zhlukov. Týmto spôsobom vznikne $n-1$ úrovní, pričom je na užívateľovi, aby rozhodol, ktorá z úrovní reprezentujúcich konkrétne zhluky predstavuje najprijateľnejšiu charakterizáciu štruktúry daných dát.

Výsledok hierarchického zhlukovania možno znázorniť pomocou dendrogramu, grafu stromovej



Obr. 1.1: Ilustračný dendrogram hierarchického zhlukovania zobrazujúci aglomeratívny smer (zdola-nahor) a opačne orientovaný divizívny smer (zhora-nadol). Ak pretneme dendrogram na vyznačenej úrovni, získame dva zhluky.

štruktúry ako možno vidieť na obrázku 1.1. Uzly na spodnej strane grafu predstavujú objekty (vstupné dáta), ktoré tvoria vlastné zhluky. Každá ďalšia úroveň predstavuje novú formáciu zhlukov, pričom možno priamo sledovať vývoj, aké objekty a v ktorej fáze do zhluku pribudnú respektíve sú z neho vylúčené. Najvrchnejší uzol reprezentuje jediný zhluk, v ktorom sa nachádzajú všetky objekty. Pretnutím dendrogramu horizontálne na požadovanej úrovni obdržíme konečný výsledok. Užívateľ v tomto kroku vyberie vhodný počet adekvátne odpovedajúci štruktúre dát. Toto vyobrazenie dát nesie významnú informačnú hodnotu, obzvlášť v prípade, že je v dátach prítomná prirodzená hierarchia.

Aglomeratívne zhlukovanie

Aglomeratívne zhlukovanie [3], [1] začína tým, že každý z n objektov je považovaný za samostatný zhluk. V ďalších krokoch postupujeme tak, že na základe vhodného kritéria a vhodnej miery spojíme dva najpodobnejšie zhluky do jedného. Nastáva redukcia zhlukov o jeden, tvorí sa

nová úroveň, ktorá označuje počet zhlukov. Tento proces sa opakuje až dokým nevznikne jeden zhluk obsahujúci všetky objekty. Musíme podotknúť, že zlúčenie dvoch zhlukov do jedného je nezvratná operácia. Vzniknutý zhluk nemožno opäť rozdeliť, či inak opraviť. Je na užívateľovi, aby rozhodol, ktorá úroveň štruktúry vhodne reprezentuje skupiny dát. Obecný postup aglomeratívnych algoritmov znázorníme nasledujúcim pseudokódom a konkrétne kritériá popíšeme samostatne podľa [3].

Algoritmus 1: Aglomeratívne zhlukovanie [3]

Vstup: databáza $\mathbf{X}^{n\times m}$, matica nepodobností $\mathbb{D}^{n\times n}$

1 Každý objekt $\mathbf{x}_i \in \mathbf{X}, i = 1, \ldots, n$ považujeme za samostatný zhluk $\mathcal{C}_1, \ldots, \mathcal{C}_n$. Počet zhlukov v aktuálnom kroku označme z:=n;

2 **repeat**

3 Prepočítame nepodobnostnú maticu $\mathbb{D}_{\mathcal{C}}^{z\times z}$ formou nepodobností (vzdialeností) zhlukov po dvojiciach;

4 V nepodobnostnej matici hľadáme $\mathbb{D}_{\mathcal{C}}$ hľadáme dva k sebe najbližšie zhluky: $\min\limits_{\substack{h,l\in\widehat{z}\\ h\neq l}} d(\mathcal{C}_h, \mathcal{C}_l)$;

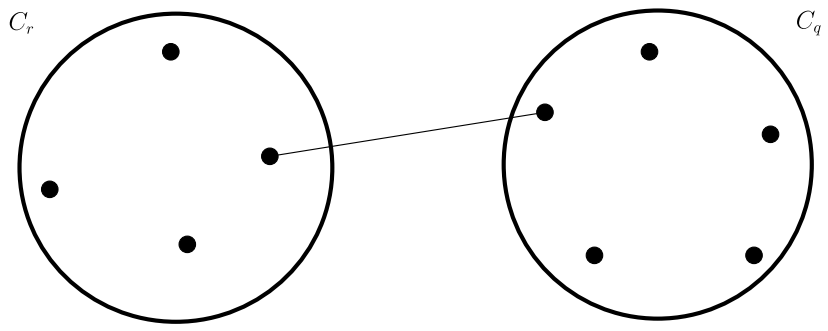
5 Zhluky s najmenšou nepodobnosťou zlúčime do jedného a z:=z-1;

6 **until** *Dostaneme jeden výsledný zhluk* $\mathcal{C}_1$*, z=1*;

Výstup: štruktúra zhlukov od n po jeden, dendrogram

Metóda najbližšieho suseda (Single linkage)

Jednou z najjednoduchších používaných metód, respektíve kritérií je metóda najbližšieho suseda [3], stretneme sa aj s prekladovým pojmom metóda jednoduchej väzby. Jej princíp spočíva v určení vzdialenosti dvoch zhlukov na báze vzdialenosti dvoch najbližších prvkov (susedov) porovnávaných zhlukov. Pre dvojicu teda platí, že ak jeden prvok patrí zhluku $\mathcal{C}_r$ druhý musí nutne patriť $\mathcal{C}_q$, túto dvojicu pre uľahčenie popisu budeme v ďalších aglomeratívnych metódach označovať ako rq-dvojica. Ak označíme vzdialenosť dvoch zhlukov $\mathcal{C}_r$ a $\mathcal{C}_q$ ako $d_{r,q}$ potom túto



Obr. 1.2: Metóda najbližšieho suseda.

vzdialenosť spočítame ako

$$d_{r,q} = \min_{\substack{\mathbf{x}_i\in\mathcal{C}_r\\ \mathbf{x}_j\in\mathcal{C}_q}} (d(\mathbf{x}_i, \mathbf{x}_j)). \tag{1.9}$$

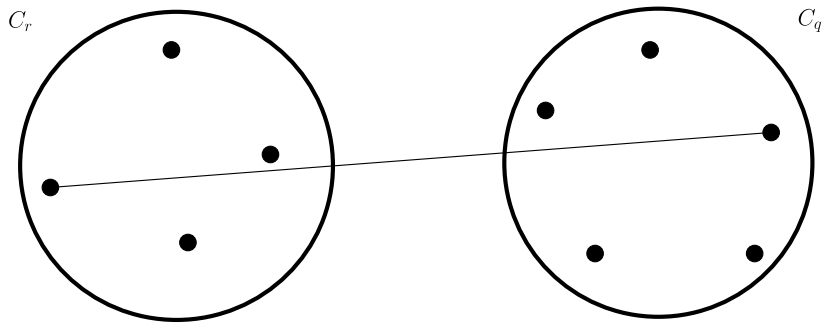
V ďalšom kroku dôjde k zlúčeniu takých dvoch zhlukov $\mathcal{C}_r$ a $\mathcal{C}_q$, pre ktoré je vzdialenosť $d_{r,q}$ minimálna spomedzi všetkých. Hľadáme teda $d = \min_{\substack{r,q \in \widehat{z} \\ r \neq q}}(d_{r,q})$.

Použitie tejto metódy sa spája s takzvaným *efektom reťazenia* [1], kedy sa prostredníctvom prechodových objektov (šumu) spoja dva zhluky, ktoré spolu často (podľa užívateľa) nesúvisia. Zhlukom vzniknutým touto metódou zväčša chýba robustnosť a sú priemerovo veľké. Voľba metódy najbližšieho suseda má výhodu v prípade, že skupiny v dátach sú rozlíšiteľné a zhluky oddelené.

Metóda najvzdialenejšieho suseda (Complete linkage)
Na opačnom extréme stavia svoje základy metóda najvzdialenejšieho suseda[1],[3]. Pri určení vzdialenosti dvoch zhlukov $\mathcal{C}_r$ a $\mathcal{C}_q$ sa zameriame na také rq-dvojice prvkov, ktoré sú si najvzdialenejšie (najnepodobnejšie). Dostávame nasledujúcu formuláciu vzdialenosti:

$$d_{r,q} = \max_{\substack{\mathbf{x}_i \in \mathcal{C}_r \\ \mathbf{x}_j \in \mathcal{C}_q}} (d(\mathbf{x}_i, \mathbf{x}_j)), \tag{1.10}$$

pričom o spojení dvoch zlukov rozhoduje, ako v predchádzajúcom prípade, najmenšia hodnota vzdialenosti zo všetkých možných dvojíc zhlukov, teda spojíme k sebe najbližšie zhluky: $d = \min_{\substack{r,q \in \widehat{z} \\ r \neq q}}(d_{r,q})$.

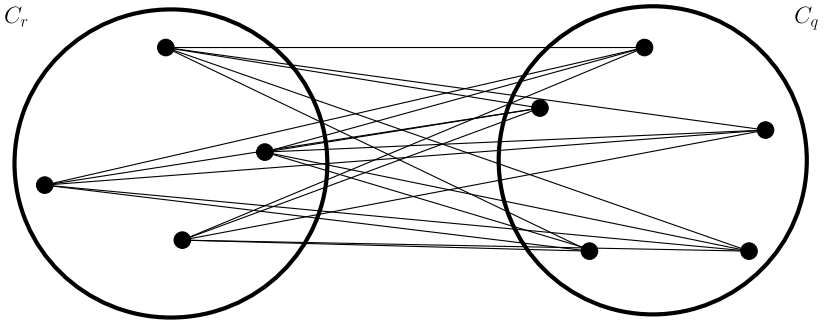


Obr. 1.3: Metóda najvzdialenejšieho suseda.

Metódu možno považovať za efektívnu z hľadiska odhalenia malých a kompaktných zhlukov. S takouto voľbou sa však spája problém, kedy prvok určitého zhluku bude bližší k prvku iného zhluku, než k prvkom zhluku vlastného [1].

Metóda priemernej väzby
Princípom metódy priemernej väzby [3] je stanovenie vzdialenosti dvoch zhlukov $\mathcal{C}_r$ a $\mathcal{C}_q$ formou priemeru vzdialností medzi všetkými rq-dvojicami. Ak označíme počet prvkov zhluku $\mathcal{C}_r$ ako n_r a počet prvkov zhluku $\mathcal{C}_q$ ako n_q, potom pre vzdialenosť týchto zhlukov platí:

$$d_{rq} = \frac{1}{n_r n_q} \sum_{\mathbf{x}_i \in \mathcal{C}_r} \sum_{\mathbf{x}_j \in \mathcal{C}_q} d(\mathbf{x}_i, \mathbf{x}_j). \tag{1.11}$$

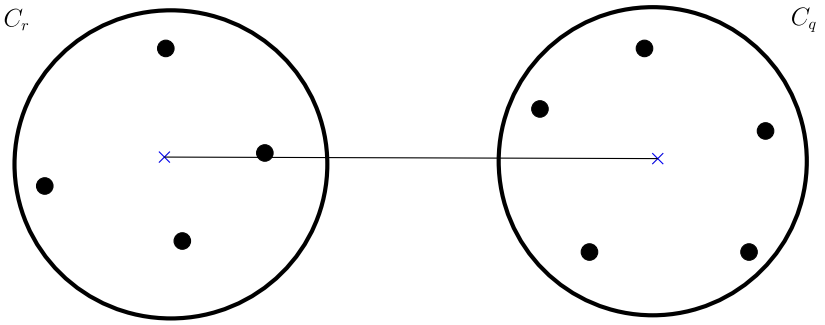


Obr. 1.4: Metóda priemernej väzby.

O spojení zhlukov opäť rozhoduje najmenšia vzdialenosť v danom kroku procesu. Výsledkom metódy sú relatívne kompaktné zhluky, dostatočne od seba vzdialené, pričom veľkým vplyvom je zvolená metrika pre vzdialenosť rq-dvojíc či numerický rozsah hodnôt premenných.

Centroidová metóda

Táto metóda [3], [2] si zakladá na práci s centroidmi. Centroidom zhluku $\mathcal{C}_h$ rozumieme priemerný objekt zhluku definovaný ako $\mathbf{c}_h = \frac{1}{n_h} \sum_{\mathbf{x} \in \mathcal{C}_h} \mathbf{x}$, kde n_h je počet objektov daného zhluku. V každej



Obr. 1.5: Centroidová metóda.

úrovni algoritmu dochádza k prepočítaniu centroidu každého zhluku. Vzdialenosť medzi dvomi zhlukmi $\mathcal{C}_r$ a $\mathcal{C}_q$ vypočítame pomocou vzdialenosti ich centier, pričom najčastejšie používame euklidovskú vzdialenosť (1.3):

$$d_{r,q} = d_E(\mathbf{c}_r, \mathbf{c}_q). \tag{1.12}$$

Rovnako ako v predchádzajúcich metódach spájame najbližšie zhluky.

Wardova metóda

Iný pohľad na tvorbu zhlukov nám poskytuje Wardova metóda [14], ktorá je často využívaná pre dáta popísané metrickým priestorom. O spojení dvoch zhlukov $\mathcal{C}_r$ a $\mathcal{C}_q$ rozhodujeme na

základe stratovej funkcie

$$J = \sum_{h=1}^{z} \sum_{\mathbf{x}_i \in \mathcal{C}_h} d_E(\mathbf{x}_i, \mathbf{c}_h), \tag{1.13}$$

kde z značí počet zhlukov v danom kroku a $\mathbf{c}_h$ je centroid h-teho zhluku. Podľa Wardovej metódy je vzdialenosť dvoch zhlukov chápaná ako prírastok stratovej funkcie po spojení dvoch zhlukov. Tvar vzdialenosti je nasledovný [2]

$$d_{r,q} = \Delta J_{r,q} = \frac{2n_r n_q}{n_r + n_q} d_E(\mathbf{c}_r, \mathbf{c}_q), \tag{1.14}$$

kde n_r a n_q sú počty objektov v zhlukoch $\mathcal{C}_r$ a $\mathcal{C}_q$. Cieľom je tento prírastok minimalizovať, pričom na počiatku, kedy každý prvok stojí samostatne, je hodnota stratovej funkcie nulová. Po nájdení takých dvoch zhlukov, ktorých prírastok je najmenší, dôjde k ich spojeniu a počet zhlukov klesne o jeden. Hodnota stratovej funkcie získaná pre pevne stanovené k zhlukov je zvyčajne vyššia, než je jej minimum pre dané k. Stretávame sa preto s riešením, ktoré spája Wardovu metódu a metódu k-means, popísanú v kapitole 2, pretože algoritmus k-means poskytuje nižšiu hodnotu stratovej funkcie (WCSS (2.2)). Kombinácia týchto metód spočíva v tom, že riešenie získané z Wardovej metódy pre k zhlukov zvolíme ako inicializačné rozdelenie pre metódu k-means. Tento spôsob zaistí zmenšenie stratovej funkcie J.

Divizívne zhlukovanie

Narozdiel od aglomeratíneho zhlukovania prebieha divizívne opačným smerom: začína stavom, kedy sú všetky objekty obsiahnuté v jedinom zhluku a rozkladovým procesom pokračuje až do fázy, kedy každý objekt stojí ako samostatný zhluk. Princíp divizívneho zhlukovania spočíva v tom, že v každej iterácií sa rozdelí najväčší zhluk (resp. zhluk s najväčšou vnútornou nepodobnosťou) na dva dcérske. Týmto spôsobom dospejeme až k stavu, kedy sa jeden z $n-1$ zhlukov rozdelí a vznikne n zhlukov po jednom objekte. Opäť je na užívateľovi aby zvážil, ktorá úroveň hierarchie je najvýstižnejšia pre dané dáta.

Kaufman a Rousseeuw [2] uvádzajú, že divizívne zhlukovanie je výpočetne náročnejšie a zriedkavejšie implementované, preto v tejto práci popíšeme len obecný algoritmus, pričom podľa autorov je možné na rovnaké dáta aplikovať ako aglomeratívne, tak aj divizívne zhlukovanie.

Aby sme boli schopní posúdiť, ktorý zhluk je nutné rozdeliť, zavedieme si podľa programu DIANA [2] veličinu priemer zhluku $\mathcal{C}_h$

$$diam(\mathcal{C}_h) = \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_h} d(\mathbf{x}_i, \mathbf{x}_j). \tag{1.15}$$

V každej iterácií rozdelíme práve jeden zhluk taký, ktorý má najväčší priemer, na dva dcérske. Použitie kritéria na základe najväčšieho priemeru možno nahradiť kritériom najväčšej vnútornej nepodobnosti. V každom kroku teda rozdeľujeme práve taký zhluk, ktorého priemerná nepodobnosť jeho členov k sebe navzájom je najväčšia. Veličina nepodobnosti zhluku $\mathcal{C}_h$ je definovaná nasledujúcim spôsobom:

$$\overline{d_{\mathcal{C}_h}} = \frac{1}{n_h} \sum_{\mathbf{x}_i \in \mathcal{C}_h} \sum_{\mathbf{x}_j \in \mathcal{C}_h} d(\mathbf{x}_i, \mathbf{x}_j), \tag{1.16}$$

kde n_h označuje počet objektov zhluku $\mathcal{C}_h$.

Algoritmus 2: Divizívne zhlukovanie

Vstup: databáza $\mathbf{X}^{n\times m}$, matica nepodobností $\mathbb{D}^{n\times n}$

Všetky objekty databázy: $\mathbf{x}_1, \ldots, \mathbf{x}_n$ vložíme do jedného zhluku $\mathcal{C}_1$.
Počet zhlukov je v tomto kroku z=1;
Načítame (resp. vypočítame) maticu nepodobností (vzdialeností) $\mathbb{D} = [d(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^{n}$;
repeat
  Zvolíme zhluk, ktorý budeme rozdeľovať (*zhluk rozdeľovaný*), ako :
    a) zhluk s najväčším priemerom $\max_{h\in\widehat{z}} diam(\mathcal{C}_h^R)$ (1.15)
    b) zhluk s najväčšou priemernou nepodobnosťou $\max_{h\in\widehat{z}} \overline{d_{\mathcal{C}_h^R}}$ (1.16);
  Vo zvolenom zhluku $\mathcal{C}_h^R$ nájdeme také dva objekty, ktoré sú od seba najviac vzdialené: $\{\mathbf{x}_i, \mathbf{x}_j | \max_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{C}_h^R} d(\mathbf{x}_i, \mathbf{x}_j)\}$. Tieto prvky budú predstavovať centrá nových zhlukov ďalšej úrovne $\mathcal{C}_h$ a $\mathcal{C}_{z+1}$;
  Ostatné prvky zhluku $\mathcal{C}_h^R$: $\mathbf{x}_h \neq \mathbf{x}_i \neq \mathbf{x}_j$ zaradíme do nových zhlukov podľa toho, ku ktorému centru ležia bližšie. Bez ujmy na všeobecnosti zvoľme $\mathbf{x}_i$ za centrum zhluku $\mathcal{C}_h$ a $\mathbf{x}_j$ ako reprezentanta zhluku $\mathcal{C}_{z+1}$;
  if $d(\mathbf{x}_h, \mathbf{x}_i) < d(\mathbf{x}_h, \mathbf{x}_j)$ **then**
    $\mathbf{x}_h$ zaradíme do $\mathcal{C}_h$;
  else
    $\mathbf{x}_h$ zaradíme do $\mathcal{C}_{z+1}$;
  $z := z + 1$;
until *Všetky objekty stoja ako samostatné zhluky* $\mathcal{C}_1, \ldots, \mathcal{C}_n$;
Výstup: štruktúra zhlukov od jedného po n, dendrogram

Stretneme sa aj s aplikáciami, kedy delenie uskutočňujeme algoritmom k-means respektíve k-medoids pre k=2, ktoré popíšeme v kapitole 2.

1.3.2 Nehierarchické zhlukovanie

Nehierarchické zhlukovanie, známe aj pod názvom rozkladové zhlukovanie (z angl. *patritioning clusthering*), je charakteristické tým, že v dátach hľadá predom zadaný počet zhlukov k. Tento proces teda, narozdiel od hierarchického zhlukovania, netvorí štruktúru rôznych počtov zhlukov. Výsledkom je práve k zhlukov, pre ktoré musí platiť, že žiaden zhluk nie je prázdny a zároveň každý objekt náleží práve jednému zhluku.

Cieľom je vytvoriť dostatočne vzdialené zhluky, v ktorých ležia sebe podobné objekty. Ide však o pomerne náročný výpočetný proces, ktorého náročnosť pre k zhlukov a n objektov je daný počtom možných rozdelení

$$P(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}C_k^i i^n. \tag{1.17}$$

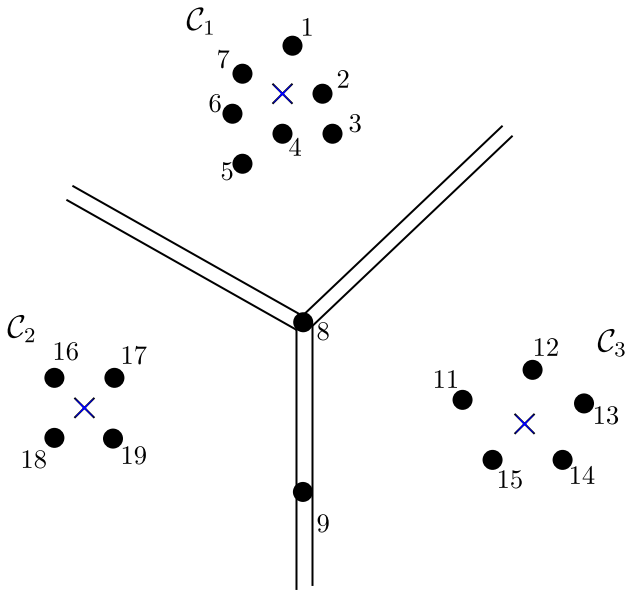
Napríklad pre 30 objektov, ktoré chceme rozdeliť do 3 skupín (zhlukov) je počet rozdelení približne $2 \cdot 10^{14}$ [3].

Nehierarchické metódy možno rozdeliť na niekoľko typov: Jednoprechodové metódy (Single-pass), relokačné metódy (Relocation) a metódy módov.
Jednoprechodové metódy sú charakteristické tým, že súbor dát sa spracúvava jedenkrát, jedným prechodom. Súbor na základe určitého kritéria roztriedime v jedinej iterácii. Metóda sa

uplatňuje najmä pre zhlukovanie dokumentov. Keďže naším cieľom je výstižne odhaliť a popísať štruktúru spoločnosti Starého Egypta, ďalej sa tejto metóde nebudeme venovať.
Relokačné metódy nájdeme aj pod názvom metódy optimalizačné. Snažíme sa teda v niekoľkých iteráciách nájsť optimálne rozloženie skupín, štruktúru ukrytú v dátach. Presúvame objekty medzi zhlukmi tak, aby sme dosiahli optimum. Hľadáme teda extrém vhodne zvolenej charakteristickej funkcie, ktorá môže popisovať napríklad *energiu väzieb zhlukov* alebo *energiu väzieb medzi prvkami rovnakého zhluku*. Najpoužívanejšou relokačnou metódou je metóda k-means (metóda k-priemerov). Metódu k-means a jej modifikácie popíšeme detailne v kapitole 2.
Poslednou kategóriou sú metódy módov a histogramov. Ich stavebným kameňom sú dáta nominálneho typu, to znamená, že každá z čŕt popisujúcich objekty nadobúda len konečný počet hodnôt. Algoritmus k-modes (k-módov) si popíšeme detailne v podkapitole 2.5.

1.3.3 Fuzzy zhluková analýza

Samostatnú skupinu zhlukovania tvorí takzvané fuzzy zhlukovanie, iným názvom prekrývajúce sa zhlukovanie. Často sa stretávame s priradením fuzzy zhlukovania ako subkategóriu nehierarchického, avšak je nutné poznamenať, že fuzzy zhlukovanie nespĺňa podmienku jedinečnosti zaradenia. Narozdiel od nehierarchického či hierarchického zhlukovania fuzzy analýza nepriraďuje objekt čisto jednému zhluku, ale určuje jeho prítomnosť vo viacerých zhlukoch. Dosiahneme teda zhluky, ktoré sa môžu vzájomne prekrývať. Kvôli absencii tejto podmienky, (*každý objekt náleží práve jednému zhluku*), v tejto práci popisujeme fuzzy analýzu zvlášť. Poznamenajme, že náš popis fuzzy analýzy v tejto práci je obecným náhľadom do problematiky, pretože fuzzy analýza je omnoho rozsiahlejšia.



Obr. 1.6: Ilustračný obrázok fuzzy zhlukovania do 3 zhlukov. Objekty 8 a 9 podľa ľudského pohľadu nepatria k ani jednému zhluku. Klasické pevné zhlukovanie by natvrdo zaradilo objekt 9 do zhluku C_2, ale fuzzy zhluková analýza určí, že objekt 9 patrí na 22% zhluku $\mathcal{C}_1$ a zároveň je uprostred zhlukov $\mathcal{C}_2$ a $\mathcal{C}_3$, pre ktoré je jeho členstvo 39% k obom rovnaké. Objekt 8 je v pevnom zhlukovaní ešte problematickejší. Leží rovnako vzdialene ku všetkým trom zhlukom a podľa fuzzy zhlukovania teda leží na približne 33% v každom zo zhlukov.

Klasické zhlukovanie, kedy objekt patrí jedinému zhluku, nazývame pevným zhlukovaním (z angl. *hard clustering*). Fuzzy zhlukovanie sa zameriava na problém z hľadiska miery príslušnosti, každý objekt je popísaný koeficientom príslušnosti, ktorý pre daný objekt určuje ako veľmi tento objekt spadá ku každému zo zhlukov. S každým objektom $\mathbf{x}_i$ je spojený vektor $\mathbf{u}_i = (u_{i1}, \ldots, u_{ik})$ známy ako koeficient príslušnosti, ktorý podľa [2] musí spĺňať nasledujúce podmienky:

$$1. \quad u_{ih} \in [0,1] \quad i = 1, \ldots, n \quad h = 1, \ldots, k, \tag{1.18a}$$

$$2. \quad \sum_{h=1}^{k} u_{ih} = 1 \quad i = 1, \ldots, n, \tag{1.18b}$$

u_{ih} príslušnosť objektu $\mathbf{x}_i$ do zhluku $\mathcal{C}_h$ je nezáporná, maximálne rovná 1 a pre každý objekt platí, že jeho celková príslušnosť (vyščítaná cez všetky zhluky) je konštantná, znormovaná na 1 (resp. môžeme chápať ako 100%). Príklad koeficientov príslušnosti je znázornený na obrázku 1.6.

Kaufman a Rousseeuw [2] vytvorili program FANNY, na ktorého základe si popíšeme bližšie fungovanie fuzzy zhlukového algoritmu. (Algoritmus FANNY je rozšírením algoritmu k-means do fuzzy logiky.) Základom pre fuzzy zhlukovanie je matica nepodobností $\mathbb{D} = [d(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^{n,n}$. Zaujímajú nás jednotlivé koeficienty príslušností. Algoritmus má za cieľ minimalizovať funkciu

$$f = \sum_{h=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ih}^2 u_{jh}^2 d(\mathbf{x}_i, \mathbf{x}_j)}{2 \sum_{j=1}^{n} u_{jh}^2}, \tag{1.19}$$

kde u_{ih} sú neznáme, ktoré sa usilujeme obdržať. Neznáme príslušnosti hľadáme tak, aby spĺňali podmienky (1.18a) a (1.18b). Výhodou algoritmu je, že takto získané stupne príslušnosti v sebe nesú detailnú informáciu o štruktúre v dátach. S tým sa však spájajú aj určité nevýhody, ktoré pri voľbe vhodnej zhlukovacej metódy musíme vziať do úvahy. S rastúcim počtom zhlukov rastie aj objem výstupnej matice príslušností, čo je z pamäťového i výpočetného hľadiska náročné. Ďalšou možnou nevýhodou je absencia reprezentačných prvkov pre jednotlivé zhluky. Výsledkom sú totiž miery príslušnosti, avšak ak chceme poznať typické črty jednotlivých zhlukov, musíme ich zistiť zvlášť.

Ak pre výsledné hodnoty mier príslušností platí

$$u_{i1} = u_{i2} = \cdots = u_{ik} \quad \forall i \in \widehat{n}, \tag{1.20}$$

každý prvok patrí rovnakou mierou do každého zhluku, hovoríme o *úplnom fuzzy zhlukovaní*. Špeciálny stav, kedy nastane, že každý objekt patrí len jednému zo zhlukov, tzn.

$$\exists! h \in \widehat{k} \text{ tak, že } u_{ih} = 1 \wedge \forall l \in \widehat{k} : \; l \neq h, \;\; u_{il} = 0 \quad \forall i \in \widehat{n}, \tag{1.21}$$

nazveme zhlukovanie *pevným* respektíve *disjunktným*. Napriek faktu, že fuzzy logika zväčša neposkytuje pevné riešenie, môžeme sa stretnúť s požiadavkou, aby sme na základe fuzzy riešenia vytvorili riešenie pevné. Za týmto účelom zavedieme nové premenné, príslušnosti pevného zhlukovania w_{ih}, ktoré nahrádzajú pôvodné príslušnosti u_{ih}. Zároveň dochádza k transformácii podmienok (1.18), aby vystihovali pevné zhlukovanie:

$$1. \quad w_{ih} = 1 \quad \vee \quad w_{ih} = 0 \quad i = 1, \ldots, n \quad h = 1, \ldots, k, \tag{1.22a}$$

$$2. \quad \sum_{h=1}^{k} w_{ih} = 1 \quad i = 1, \ldots, n, \tag{1.22b}$$

ktoré hovoria, že objekt $\mathbf{x}_i$ buď patrí, alebo nepatrí zhluku $\mathcal{C}_h$ a že každý objekt možno zaradiť práve jedenkrát. Objektu $\mathbf{x}_i$ stanovíme $w_{iq} = 1$ pre taký zhluk $\mathcal{C}_q$, pre ktorý koeficient u_{iq} nadobúda najvyššiu hodnotu. Táto transformácia však nevylučuje také špeciálne prípady, kedy môže nastať, že niektoré pevné zhluky zostanú prázdne. Aby sme dokázali povedať, ako veľmi sa líši fuzzy zhlukovanie od pevného, čiže ohodnotiť fuzzy zhlukovanie, môžeme použiť *Dunnov koeficient rozkladu* [2], definovaný ako podiel sumy štvorcov všetkých koeficientov príslušností a počtu objektov

$$F_k(\mathbb{U}) = \sum_{i=1}^{n}\sum_{h=1}^{k}\frac{u_{ih}^2}{n}, \tag{1.23}$$

kde $\mathbb{U}$ je matica všetkých koeficientov príslušností

$$\mathbb{U} = [u_{ih}]_{n\times k} = \begin{pmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,k} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n,1} & u_{n,2} & \cdots & u_{n,k} \end{pmatrix}.$$

Hodnota Dunnovho koeficientu rozkladu dosahuje maximálne hodnotu 1, kedy dochádza k pevnému zhlukovaniu, a naopak nadobúda minimum $\frac{1}{k}$, kedy všetky príslušnosti nadobúdajú hodnotu $\frac{1}{k}$. Tieto prípady sú ukázané v nasledujúcich dvoch bodoch:

- $\forall i \in \widehat{n} \quad \exists l \in \widehat{k},\ \ u_{il} = 1\ \ u_{ih} = 0$ pre $l \neq h$

$$F_k(\mathbb{U}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{k}u_{ih}^2 = 0 + \frac{1}{n}\sum_{i=1}^{n}(u_{il})^2 = \frac{n}{n} = 1,$$

- $\forall i \in \widehat{n} \quad \forall h \in \widehat{k}: \quad u_{ih} = \frac{1}{k}$

$$F_k(\mathbb{U}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{k}u_{ih}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{k}\frac{1}{k}^2 = \frac{1}{n}nk\frac{1}{k^2} = \frac{1}{k}.$$

Tento koeficient možno normalizovať tak, aby nadobúdal hodnoty z intervalu $[0, 1]$:

$$F_k'(\mathbb{U}) = \frac{F_k(\mathbb{U}) - \frac{1}{k}}{1 - \frac{1}{k}} = \frac{kF_k(\mathbb{U}) - 1}{k - 1}. \tag{1.24}$$

Okrem výstupu vo forme tabuľky príslušností objektov k zhlukom je vhodným grafickým výstupom fuzzy zhlukovania aj takzvaný obrysový graf [4]. Tento graf popisuje vzťah jednoznačného zaradenia objektu a možnosti vytvorenia prekrývajúcich sa zhlukov. Poznamenajme, že ide o nástroj, ktorý nie je výlučne viazaný na fuzzy analýzu, ale jeho použitie je mnohostranné. Bližší popis a použitie uvádzame v kapitolách 3 a 4.

Kapitola 2

Metóda K-means: princípy a modifikácie

Jednou z najznámejších a najpoužívanejších metód nehierarchického zhlukovania je metóda k-means (metóda k-priemerov). Radí sa medzi ľahko použiteľné metódy a často sa stretávame s použitím na už predspracované dáta. Algoritmus je teda možno kombinovať aj s iným typom metód, kedy vznikajú napríklad kombinácie hierarchického zhlukovania a algoritmu k-means. Disjunktné zhluky dát podobne ako k-means vytvára aj jeho pozmenená varianta metóda k-medoids. V tejto kapitole popíšeme k-means a jeho variácie k-means++, k-medoids, dvojfázový k-means, k-modes, pričom vychádzame najmä z [2], [3], [10], [6], [9] a [11].

2.1 K-means

Prvýkrát k-means predstavil MacQueen v roku 1967 [5]. Ide o univerzálnu metódu s cieľom rozklasifikovať pozorované dáta do k zhlukov. Každý zhluk obsahuje také objekty, ktoré sú si vnútri zhluku podobnejšie než s objektmi z iných zoskupených zhlukov. Každý zhluk nesie charakteristické hodnoty príznakov, podľa ktorých možno popísať každý jeho objekt. Vektory týchto hodnôt príznakov nazývame centroidmi. Pre konkrétny zhluk, ktorý označíme C_h, kde $h \in \widehat{k}$, označíme centroid, teda vektor príznakov ako $\mathbf{c}_h$. Centroid je definovaný ako vektor, pre ktorý je súčet vzdialeností všetkých objektov k nemu samému minimálny. Inými slovami, centroid predstavuje ťažisko všetkých objektov náležiacich zhluku. Ak pracujeme s Euklidovskou vzdialenosťou (1.3) vďaka čomu centroid je priemerom:

$$\mathbf{c}_h = \frac{1}{|C_h|} \sum_{\mathbf{x}_i \in C_h} \mathbf{x}_i \quad pre \quad h \in \widehat{k} \quad a \quad i \in \widehat{n}, \tag{2.1}$$

hovoríme o metóde k-priemerov.

Majme $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ dátovú množinu n objektov a nech k je prirodzené číslo. K-means algoritmus hľadá k vektorov $\mathbf{c}_h$, $h = 1, \ldots, k$ tak, aby sa minimalizovala *vnútrozhluková suma štvorcov WCSS* (z angl. *Within Cluster Sum of Squares*)

$$WCSS = \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} d_E^2(\mathbf{x}_i, \mathbf{c}_h). \tag{2.2}$$

$$J = WCSS = \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} d_E^2(\mathbf{x}_i, \mathbf{c}_h)$$
$$= \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} \sum_{j=1}^{m} (\mathbf{x}_{ij} - \mathbf{c}_{hj})^2.$$

Aby sme zistili, kde leží centrum, ktoré minimalizuje vnútrozhlukovú energiu J, zderivujeme predpis po zložkách a položíme rovno 0.

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{c}_{rl}} &= \frac{\partial}{\partial \mathbf{c}_{hj}} \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} \sum_{j=1}^{m} (\mathbf{x}_{ij} - \mathbf{c}_{hj})^2 \quad \forall r \in \widehat{k}, \forall l \in \widehat{m} \\
&= \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} \sum_{j=1}^{m} \frac{\partial}{\partial \mathbf{c}_{rl}} (\mathbf{x}_{ij} - \mathbf{c}_{hj})^2 \\
&= \sum_{\mathbf{x}_i \in C_r} 2(\mathbf{x}_{ij} - \mathbf{c}_{rl}) \\
0 &\stackrel{!}{=} 2 \sum_{\mathbf{x}_i \in C_r} (\mathbf{x}_{ij} - \mathbf{c}_{rl}) \\
0 &\stackrel{!}{=} \sum_{\mathbf{x}_i \in C_r} \mathbf{x}_{ij} - \sum_{\mathbf{x}_i \in C_r} \mathbf{c}_{rl},
\end{aligned} \tag{2.3}$$

ak označíme n_r ako počet objektov v zhluku C_r, dostávame:

$$n_r \mathbf{c}_{rl} = \sum_{\mathbf{x}_i \in C_r} \mathbf{x}_{ij},$$

čím po predelení rovnice získame tvar zložky centra

$$\mathbf{c}_{rl} = \frac{1}{n_r} \sum_{\mathbf{x}_i \in C_r} \mathbf{x}_{ij} = \overline{\mathbf{x}_{ij}} \quad \forall r \in \widehat{k}, \forall l \in \widehat{m}. \tag{2.4}$$

A môžeme písať $\mathbf{c}_r = \overline{\mathbf{x}_i}$ pre $\forall r \in \widehat{k}$. To znamená, že centroid v tvare priemeru prvkov v zhluku minimalizuje WCSS, čo predstavuje hlavný pilier metódy k-means a jej adaptácií.

Principiálne v prípade k-means hovoríme o dvojkrokovej iteratívnej procedúre. Začína definovaním k inicializačných centroidov, kde každý reprezentuje iný zhluk. Voľba týchto centroidov je možná rôznymi spôsobmi. Môžeme vybrať náhodne k objektov, respektíve prvých k objektov a následne ich prehlásiť za centroidy. Keď už disponujeme centroidmi, každému centroidu priradíme také objekty, ktorých vzdialenosť k nim je najmenšia v porovnaní so vzdialenosťou k ostatným potencionálnym centrám zhlukov. Následne vypočítame polohu nového centra ako vektor priemerov znakov jednotlivých objektov patriacich do rovnakého zhluku. Každý krok má za cieľ minimalizovať $WCSS$ až konverguje k lokálnemu minimu. Kroky priraďovania prvkov k centroidom a ich prepočítavania sa opakujú až kým nie je zaradenie konštantné a centroidy sa viac nemenia.
Nasledujúci popis algoritmu sme prevzali z [6]:

Algoritmus 3: K-means [6]

Vstup: $k \in \mathbf{N}$, databáza $\mathbf{X}^{n \times m}$

1 Voľba inicializačných centroidov $\mathbf{c}_1, \ldots, \mathbf{c}_k$;
2 **repeat**
3 Zaradíme každý z objektov do zhluku najbližšieho (najpodobnejšieho) centroidu danému objektu;
4 Prepočítame centroidy ako vektory priemerov čŕt zhluku náležiacich objektov;
5 **until** *poloha centroidov je konštantná*;

Výstup: Množina k zhlukov, ktoré minimalizujú kvadratické kritérium 2.2

Je nutné poznamenať, že počiatočná voľba centroidov významne ovplyvňuje algoritmus. Typicky ak na začiatku procesu nemáme znalosť resp. postačujúci odhad o polohe centroidov, algoritmus môžeme spustiť niekoľkokrát pre rôzne náhodné inicializačné varianty. Výsledky niekoľkých realizácii porovnáme a ponecháme najlepšie možné riešenie. Obdobne aj voľba hyperparametru k nemusí byť jednoznačná a v procese hľadáme odhad poskytujúci vhodné riešenie. Algoritmus realizujeme pre rôzne vstupné hodnoty k a následne porovnáme výsledky. K výberu najadekvátnejšej hodnoty k môžeme použiť napríklad *Elbow method, The Silhouette a pod.* ktoré bližšie popíšeme v časti 3.

2.2 K-means++

V tejto podkapitole sa budeme zaoberať modifikáciou metódy k-means, takzvanou metódou k-means++. Cieľom novej modifikácie je zvýšiť efektivitu a presnosť pôvodnej metódy. K-means++ je založený na kombinácii špecifickej voľby inicializačných centroidov a pôvodného algoritmu k-means. Práve táto kombinácia podľla Davida Arthura a Sergeia Vassilvitského [9] zaisťuje optimalizačne presnejšie a rýchlejšie riešenie s výpočetnou náročnosťou rádovo O(log$k$)

Algoritmus 4: K-means++ [9]

Vstup: $k \in \mathbf{N}$, databáza $\mathbf{X}^{n \times m}$

1 Voľba inicializačných centroidov $\mathbf{c}_1, \ldots, \mathbf{c}_k$ podľa algoritmu 5;
2 Algoritmus klasického k-means bez prvého kroku voľby inicializačných centier;

Výstup: Množina k zhlukov, ktoré minimalizujú kvadratické kritérium (2.2)

Kompozícia algoritmu nám umožňuje zameniť v druhom bode k-means za inú jeho modifikáciu. Po takejto zmene nám však článok nič negarantuje.

Pozrieme sa na spôsob inicializácie centier zhlukov prostredníctvom algoritmu 5. Označme $D(x)$ ako najkratšiu vzdialenosť objektu x od jeho najbližšieho centra, ktoré už bolo zvolené. Pri výpočte používame euklidovskú vzdialenosť (1.3) medzi dvojicou objektov. Tento pseudokód inicializácie centier 5 sa zakladá na pravdepodobnostnom výbere. Nové centrum sa vyberá na základe pravdepodobnostného rozdelenia (2.5) z bodov databázy $\mathbf{X}$.

Algoritmus 5: Inicializačný algoritmus pre k-means++

Vstup: $k \in \mathbf{N}$, databáza $\mathbf{X}^{n \times m}$

1 Zvolíme práve jedno centrum c_1 uniformne náhodne z $\mathbf{X}$;

2 Vyberieme nové centrum $c_i = x\prime \in \mathbf{X}$ s pravdepodobnosťou:

$$p(x\prime) = \frac{D(x\prime)^2}{\sum_{x \in \mathbf{X}} D(x)^2}; \tag{2.5}$$

3 Opakujeme pre $i = 2 \dots k$, až máme všetky centrá;

Výstup: Množina $\mathcal{C} = (\mathbf{c}_1, \dots, \mathbf{c}_k)$ inicializačných centier

Arthur a Vassilvitskii [9] experimentálne ukázali, že za určitých podmienok možno docieliť konvergenciu v super-polynomiálnom čase a dokázali nasledujúcu vlastnosť.

Veta 2.2.1. Ak množina centier $\mathcal{C}$ je skonštruovaná pomocou k-means++, potom odpovedajúca potenciálová funkcia J spĺňa

$$E[J] \leq 8(\ln k + 2) J_{\mathrm{OPT}}, \tag{2.6}$$

kde J_{OPT} značí optimálnu hodnotu potenciálovej funkcie.

Poznamenajme, že sa môžeme stretnúť aj s inými pomenovaniami potenciálovej funkcie, napr. *error function, partitioning cost, etc.*. Vedzme však, že všetky predstavujú celkovú odchýlku objektov od svojich centier, teda hodnotu potenciálu nad nulovou hladinou, ktorú predstavujú centroidy.

2.3 K-medoids

Formálne aplikácia metódy k-means závisí na podmienke použitia takej miery nepodobnosti pre dvojicu objektov, ktorá je ekvivalentná ich Euklidovskej vzdialenosti. To znamená, že sa obmedzujeme na kvantitatívny typ premenných reprezentujúcich črty skúmaných dát. Práve použitie Euklidovskej vzdialenosti spôsobuje, že k-means nie je odolný voči odľahlým objektom, tzv. *outliers*. Odľahlé objekty majú za následok vysoké hodnoty v nepodobnostnej miere, inými slovami, ich vzdialenosť od ostatných objektov je veľmi vysoká. Tento fakt (nepriaznivo) ovplyvňuje kritérium (2.2).

Ako už sme spomínali v úvode kapitoly 1, cieľom zhlukovej analýzy je nájsť spôsob, akým môžeme charakterizovať sebe podobné skupiny v súbore dát. Príslušnú charakterizáciu v sebe nesie reprezentant vlastností, v prípade algoritmu k-means ním bol vektor priemerov vlastností respektíve akési ťažisko zhluku. Avšak, toto ťažisko je fiktívny bod príznakového priestoru a teda nenáleží do našej databázy $\mathbf{X}$, pričom sa môžeme stretnúť s takou požiadavkou úlohy zhlukovania, aby reprezentant zhluku bol práve objekt databázy $\mathbf{X}$. Splnenie takejto podmienky nám poskytuje algoritmus k-medoids. Definujeme totiž medoid, objekt zhluku, pre ktorý je priemerná nepodobnosť k ostatným objektom toho samého zhluku minimálna. Stretneme sa aj s označením PAM *(Partitioning around medoids)* podľa Kaufmana a Rousseeuwa [2]. V tejto časti práce primárne využívame popis metódy podľa Wierzchoń a Kłopotka [10].

K-medoids algoritmus hľadá k centier zhlukov, takzvaných medoidov, ktoré sú objektmi náležiacimi databáze $\mathbf{X}$ a zároveň pre ne funkcia

$$J_{med} = WCSS_{med} = \min_{1 \leq h \leq k} \sum_{x_i \in C_h} d(\mathbf{x}_i, \mathbf{c}_h), \tag{2.7}$$

dosiahne minimum.
Poznamenajme, že pre použitú mieru nepodobnosti $d : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ nemusí platiť symetria či nemusí ísť vyslovene o metriku (narozdiel od Euklidovskej vzdialenosti).
Výhodou metódy je taktiež možnosť voľby vstupných dát buď vo forme matice vektorov objektov, alebo vo forme nepodobnostnej matice, ktorá je dimenzionálne závislá len na počte objektov.
Pri voľbe nepodobnostnej matice už reprezentácia objektu v priestore čŕt nehrá úlohu a praktická implementácia algoritmu využíva len vektor $\mathbf{c}$ s elementami c_i, ktoré indikujú, akému zhluku i-ty objekt patrí:

$$c_i = \begin{cases} h & \text{ak } x_i \in C_h, \\ i & \text{ak } x_i \text{ je medoid (exemplár).} \end{cases} \tag{2.8}$$

Pseudokód algoritmu k-medoids preberieme z [10].

Algoritmus 6: K-medoids [10]

Vstup: $k \in \mathbf{N}$, nepodobnostná matica $\mathbf{D} = [d_{ij}]_{n \times n}$

1 Výber podmnožiny $\mathcal{K} \subset \{1, \ldots, n\}$, ktorá značí množinu medoidov;
2 **while** *mení sa zaradenie objektov* **do**
3 Zaradíme každý z objektov do zhluku použitím pravidla

$$c_i = \begin{cases} \arg\min\limits_{h \in \mathcal{K}} d_{ih} & \text{ak } i \notin \mathcal{K}, \\ i & \text{inak} \end{cases} \tag{2.9}$$

pre $i = 1, \ldots, n$;
4 Zaktualizujeme medoidy spôsobom:

$$h_r^* = \arg\min_{t : c_t = r} \sum_{t' : c_{t'} = r} d_{tt'}, \tag{2.10}$$

kde $r = 1, \ldots, k$;

Výstup: Množina zaradenia $\mathcal{C} = \{C_1, \ldots, C_k\}$

Inicializácia počiatočných medoidov $\mathcal{K}$ môže prebehnúť náhodne alebo častejšie voľbou k tých najnepodobnejších objektov k ostatným. Podmienka (2.9) určuje zaradenie objektu i k zhluku takého medoidu, s ktorým má najmenšiu hodnotu nepodobnostnej funkcie, inými slovalmi, vzdialenosť medzi nimi je najkratšia. A podmienka (2.10) hovorí, že pre objekty blízke spoločnému medoidu (prvky patriace k r-tému medoidu), zvolíme nový tak, aby suma nepodobností k ostatným objektom tohto zhluku bola najnižšia možná. Celý algoritmus končí vo chvíli, keď medoidy zostávajú fixné, teda ako v prípade k-means, centrá sa nemenia.

2.4 Dvojfázový k-means

Pomocou metód zhlukovej analýzy sa snažíme vytvoriť a popísať model napríklad marketingovej stratégie alebo model umožňujúci náhľad so štruktúry populácie (väzby medzi poddruhmi fauny a flóry či v našom prípade charakter a vývoj spoločnosti Starého Egypta). V praxi sa v dátach, určených k analýze, vyskytujú odľahlé prvky, ktoré majú značný vplyv na tvorbu zhlukov a ich výslednú charakterizáciu. Rozumieme tým, že vzdialená poloha odľahlých objektov v príznakovom priestore spôsobuje vychýlenie centra zhluku, do ktorého boli zaradené a

tým ovplyvnia vnútrozhlukovú, či mimozhlukovú vzdialenosť. Ich identifikácia je preto dôležitá z hľadiska precíznosti modelu popisujúceho dáta.

Môžeme sa stretnúť s algoritmami, ktoré vzdialené prvky zanedbávajú alebo im priraďujú nízku váhu. Odolnosť metódy k-medoids spočíva v špecifickej voľbe inicializácie a v následnej práci s vlastnými objektmi ako centrami. V niektorých zhlukovacích problémoch potrebujeme špeciálne nájsť anomálie v dátach, teda odhaliť resp. detekovať práve odľahlý prvok. Napríklad pokiaľ máme dáta reakcií pacientov po podaní testovacích liečiv, je dôležité odhaliť práve atypické vedľajšie účinky, ktoré sa môžu správať ako prvky odľahlé.

Vhodnou aplikáciou pre túto problematiku je dvojfázový k-means, ktorý popísali Jiang, Tseng a Su vo svojom článku [11]. V tejto sekcii budeme z článku vychádzať. Autori pojmom *outlier*, teda odľahlým prvkom, rozumejú menší zhluk, ktorý je ďaleko od ostatných objektov. Majme na pamäti, že zhluk tvoria sebepodobné objekty, teda je možné aby zhluk bol tvorený jediným prvkom. Metóda sa skladá z dvoch fáz:

- modifikovaný proces metódy k-means (MKP, modified k-means process),
- proces detekcie odľahlých objektov (OFP, outliers-finding process).

V prvej fáze sa využíva modifikácia metódy k-means založená na heuristike: *ak je vkladaný objekt príliš vzdialený od ostatných centier zhlukov, potom je prehlásený za centrum nového zhluku.* To má podľa autorov za následok, že v jednom zhluku sú buď všetky objekty odľahlé alebo žiaden z objektov nie je odľahlý. Zmienime aj vedľajší efekt, a to vznik väčšieho počtu zhlukov než je k. Tento problém rieši práve druhá fáza, založená na princípe vyčleňovania odľahlých prvkov prednostne. Popíšeme si obe fázy detailne v nasledujúcich dvoch sekciách.

2.4.1 1. fáza: Modifikovaný proces metódy k-means

Rozdiel od tradičnej metódy k-means spôsobuje práve spomínaná heuristika. Spôsobuje fluktuáciu počtu vytvorených zhlukov v určitom rozmedzí. Napriek rovnakému vstupu získame iný počet zhlukov. Vznik vyššieho počtu zhlukov nám neskôr pomôže odhaliť odľahlé prvky alebo takzvaný zvyškový zhluk, to jest zhluk obsahujúci v porovnaní s ostatnými také objekty, ktoré sú si rozdielne, ale ich zaradenie k iným je nevhodné.
Začiatok modifikovaného algoritmu sa od klasického nelíši. Máme k dispozícii dátovú maticu $\mathbf{X}$ a požadovaný počet zhlukov k. Zvolíme k' regulovateľný počet zhlukov a položíme $k' := k$. Následne náhodne spomedzi objektov $\mathbf{X}$ volíme k' inicializačných centier $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{k'}\}$, $\mathbf{c}_i \in \mathbb{R}^n$, $\in \widehat{k'}$.

V priebehu každej iterácie, kedy zaradíme objekt k najbližšiemu zhluku, počítame okrem aktualizovanej polohy centier aj vzájomnú zhlukovú vzdialenosť, pričom nás zaujíma najmenšia vzdialenosť dvojice centier:

$$min(\mathcal{C}) = min(d_E(\mathbf{c}_h, \mathbf{c}_l)), \quad \text{pre} \quad h, l = 1, \ldots, k' \quad a \quad l \neq h. \tag{2.11}$$

Pre každý objekt $\mathbf{x}_i$, $i \in \widehat{n}$ je najkratšia vzdialenosť $d_{min}(\mathbf{x}_i, \mathcal{C})$ k najbližšiemu centru zhluku daná vzťahom:

$$d_{min}(\mathbf{x}_i, \mathcal{C}) = min(d_E(\mathbf{x}_i, \mathbf{c}_h)) \quad pre \quad h = 1, \ldots, k'. \tag{2.12}$$

Vďaka použitej heuristike môže nastať situácia, že počet vzniknutých zhlukov bude väčší než k. Zakaždým, keď nájdeme objekt, ktorý leží od centra ďalej než centrum iného zhluku, objekt oddelíme. Takto môže nastať situácia, že zhluky pribúdajú. V extrémnom prípade môže byť každý bod zaradený samostatne do svojho vlastného zhluku a teda $k' = n$, pričom n je počet dát.

Takémuto riešeniu sa snažíme vyhnúť a preto si stanovíme k_{max} ako maximálny možný počet zhlukov, pre ktorý má zhlukovanie ešte zmysel. Musí platiť, že $k \leq k_{max} \leq n$. Zároveň si stanovíme hraničnú podmienku pre k', tak aby sme ani regulovateľným počtom zhlukov nepresiahli maximálny zmysluplný počet: $k' \leq k_{max}$.

Algoritmus ilustrujeme pomocou pseudokódu:

Algoritmus 7: Modifikovaný algoritmus k-means heuristikou [11]

Vstup: Dátová matica $\mathbf{X}$,$k' = k \in \mathbf{N}$

1 Náhodná voľba inicializačných centier $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{k'}\}$;
2 **repeat**
3 Spočítame $min(\mathcal{C})$ a $d_{min}(\mathbf{x}_i, \mathcal{C})$ pre $i = 1, \ldots, n$;
4 Ak $d_{min}(\mathbf{x}_i, \mathcal{C}) \leq min(\mathcal{C})$ prechod na krok 7;
5 **Rozdeľovací proces:** ak $d_{min}(\mathbf{x}_i, \mathcal{C}) > min(\mathcal{C})$ pre $i \in \widehat{n}$, potom $\mathbf{x}_i$ bude centrum nového zhluku a $k' := k' + 1$;
6 **Zlučovací proces:** ak $k' > k_{max}$, potom zlúčime dva najbližšie zhluky do jedného a $k' = k_{max}$;
7 **Alokačný proces:** priradenie $\mathbf{x}_i$ najbližšiemu zhluku;
8 **until** *zaradenie zostane stabilné, tj. pozícia centier sa takmer nemení*;

Výstup: Množina centroidov $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{k'}\}$

Zlučovací proces má za úlohu spojiť dva zhluky, ktorých centrá sú si najbližšie (najpodobnejšie). Centrum novovzniknutého zhluku $\mathbf{c}_h^*$ určíme vzťahom:

$$\mathbf{c}_h^* = \frac{1}{N_{h1} + N_{h2}}[N_{h1}\mathbf{c}_{h1} + N_{h2}\mathbf{c}_{h2}], \tag{2.13}$$

kde $\mathbf{c}_{h1}$ a $\mathbf{c}_{h2}$ sú centrá najbližších zhlukov vstupujúcich do procesu a N_{h1} a N_{h2} je populácia jednotlivých zhlukov. Procesom sa zredukuje počet zhlukov o jeden. V každej iterácií pod pojmom **alokačný proces** rozumieme znovuzaradenie objektov do jednotlivých, im najbližších, zhlukov s cieľom minimalizovať vnútrozhlukovú vzdialenosť (2.2).

Dokončením tejto fázy dostaneme k' plne určených zhlukov.

2.4.2 2. fáza: Proces detekcie odľahlých objektov

Po obdržaní k' zhlukov z prvej fázy je náplňou fázy druhej detekovať odľahlé objekty a vytvoriť finálnych k zhlukov. V tejto fáze sa môžeme vydať dvoma smermi. K odhaleniu odľahlých objektov možno použiť niektorú z hierarchických metód, ktoré sme podrobnejšie popísali v časti 1.3.1 a určiť odľahlý prvok prostredníctvom vysokej vzdialenosti. Poznamenajme, že táto metóda má zložitosť $O(n^3)$. Druhou voľbou je použiť princíp *minimálnej kostry grafu*, kde sa uvádza zložitosť $O(n^2)$. Práve z dôvodu nižšej zložitosti popisujeme ďalej princíp minimálnej kostry.

K dispozícií máme k' zhlukov, ktoré budeme považovať za uzly grafu. Stačí sa teda zamerať len na centrá zhlukov. Najkratšia vzdialenosť dvoch uzlov, teda vzdialenosť dvoch centier, tvorí ohodnotenú hranu. Dostávame teda úplný ohodnotený graf. Ďalej predpokladajme, že na začiatku procesu fázy dva máme les $\mathcal{F}$, ktorý je zatiaľ prázdny. V prvom kroku skonštruujeme minimálnu kostru grafu k' uzlov. Vznikne strom, ktorý umiestnime do lesa $\mathcal{F}$. Následne odstránime najdlhšiu hranu zo stromu a nahradíme strom dvoma novovzniknutými podstromami, ktoré umiestnime do lesa $\mathcal{F}$. Stromy s menším počtom uzlov prehlásime za odľahlé objekty, resp. podľa potreby vylúčime z ďalšieho spracovania. Takýmto spôsobom opakovane odstraňujeme hrany s najvyšším ohodnotením až do doby, kedy sa v lese $\mathcal{F}$ nachádza presne k stromov. Popísaný proces zhrnieme prostredníctvom nasledujúceho pseudokódu:

Algoritmus 8: Proces detekcie odľahlých objektov [11]

Vstup: Množina zhlukov $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{k'}\}$ z fázy 1, les $\mathcal{F} = \emptyset$

1 Konštrukcia minimálnej kostry grafu z uzlov z množiny $\mathcal{C}$, vznik stromu;
2 Vložíme strom do lesa $\mathcal{F}$;
3 **repeat**
4 | Odstránime najdlhšiu hranu jednému zo stromov z lesa $\mathcal{F}$. Pôvodný strom nahradíme dvoma práve vzniknutými podstromami;
5 | Podstromy s menším počtom uzlov označíme za odľahlé (resp. odstránime);
6 **until** *počet stromov v lese* $\mathcal{F}$ *je rovný* k;

Výstup: Rozdelenie dát do k zhlukov s nájdenými odľahlými objektami

2.5 K-modes

Poslednou metódou, ktorú si predstavíme, je metóda k-modes (k-módov). Ide o adaptáciu k-means na objekty, ktorých črty sú nominálneho typu. Každá črta zvyčajne nadobúda hodnoty z konečnej množiny, napríkad jedna z čŕt objektu môže predstavovať pohlavie { žena, muž, iné }, spoločenský status {šľachta, duchovenstvo, pracujúca vrstva, poddaní, ... }, a pod.. Aby sme mohli aplikovať metódy zhlukovej analýzy, nahradíme tieto hodnoty čŕt postupnými celými číslami, podľa [10]. Dostaneme napríklad hodnoty črty pohlavia ako { 1,2,3 }. Tým sa dostávame k číselnej reprezentácii objektu $\mathbf{x}_i$ pre $\forall i \in \widehat{n}$.

Našou úlohou v zhlukovej analýze je zoskupiť sebepodobné resp. k sebe blízke objekty. Preto definujeme nasledujúcu mieru nepodobnosi objektov $\mathbf{x}_i$ a $\mathbf{x}_j$, vhodnú pre aplikáciu na nominálne dáta, ako počet rozdielov medzi objektmi:

$$d_n(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{m} \delta(x_{il}, x_{jl}),$$
$$\text{kde } \delta(x_{il}, x_{jl}) = \begin{cases} 1 & \text{ak } x_{il} \neq x_{jl}, \\ 0 & \text{inak.} \end{cases} \tag{2.14}$$

Mód množiny $\mathcal{C}$ definujeme ako objekt $\mathbf{c}^{mod}$ (nemusí byť nutne prvkom databázy $\mathbf{X}$), ktorý minimalizuje nepodobnosť množiny voči nemu:

$$d_n(\mathcal{C}, \mathbf{c}^{mod}) = \sum_{\mathbf{x} \in \mathcal{C}} d_n(\mathbf{x}, \mathbf{c}^{mod}). \tag{2.15}$$

Aby sme boli schopný takýto mód zostrojiť, potrebujeme k tomu nasledujúcu Lemmu prevzatú z [10].

Lemma 2.5.1. Nech $Dom(A_l)$ označuje množinu hodnôt črty A_l a označme počet objektov, pre ktoré v množine $\mathcal{C}$ l-tá črta nadobúda hodnotu s ako a_{ls}. Potom ak zložky vektora $\mathbf{c}^{mod}$ spĺňajú podmienku

$$\mathbf{c}_l^{mod} = \underset{s \in Dom(A_l)}{\arg\max}\ a_{ls}, \quad l = 1, \ldots, m, \tag{2.16}$$

vektor $\mathbf{c}^{mod}$ je módom množiny $\mathcal{C}$.

Proces metódy k-modes znázorníme nasledujúcim pseudokódom:

Algoritmus 9: k-modes [10]

Vstup: Databáza $\mathbf{X}$, počet zhlukov k

1 Zvoľme ľubovoľne k inicializačných módov;
2 **repeat**
3 Podľa miery nepodobnosti d_n (2.14) zaradíme objekty k najbližším módom - do najbližšieho zhluku;
4 Podľa lemmy 2.5.1 aktualizujeme mód pre každý zhluk;
5 **until** *Rozdelenie a poloha módov zostávajú nemenné*;

Výstup: Rozdelenie dát do zhlukov $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$

Úlohu k-modes možno popísať aj pomocou indexu vnútrozhlukovej vzdialenosti a to tak, že metóda má za cieľ index

$$J_{mod}(\mathbf{c}_1^{mod}, \ldots, \mathbf{c}_k^{mod}) = \sum_{h=1}^{k} d_n(\mathcal{C}_h, \mathbf{c}_h^{mod}) = \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{C}_h} d_n(\mathbf{x}_i, \mathbf{c}_h^{mod}) \tag{2.17}$$

minimalizovať.

2.5.1 Metóda k-prototypes

Metódu k-modes možno upraviť pre zhlukovací problém, v ktorom vystupujú črty dvoch typov: nominálneho (pohlavie, spoločenský status...) a merateľného typu (vek, výška, váha...). Pre tento problém upravíme usporiadanie čŕt tak, aby prvých n_1 zložiek vektora objektu boli črty merateľného typu a zvyšné budú nominálne.
Označme maticu reprezentujúcu zaradenie objektov do zhlukov ako $\mathbb{U} = [u_{ih}]$, rozmeru $n \times k$, tak že $u_{ih} = 1$ ak $\mathbf{x}_i \in \mathcal{C}_h$ a $u_{ih} = 0$ inak. Označme ďalej množinu centier, (módov/prototypov) ako $\mathbf{M} = \{\mathbf{c}_1^{mod}, \ldots, \mathbf{c}_k^{mod}\}$. Potom rozložíme spôsob počítania miery nepodobnosti na dve časti podľa typu črty, konkrétne pre merateľné črty použijeme druhú mocninu Euklidovskej vzdialenosti d_E (1.3) a pre nominálne črty použijeme nominálnu nepodobnosť d_n (2.14). Celková nepodobnosť (nepodobnosť pre prototypy) má teda tvar:

$$d_n(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{n_1} (x_{il} - x_{jl})^2 + \gamma \sum_{l=n_1+1}^{m} \delta(x_{il}, x_{jl}), \tag{2.18}$$

kde $\gamma > 0$ je balančný koeficient, vďaka ktorému sa vyhneme uprednostneniu jedného typu premenných.

S takto zavedenou mierou môžeme obdobne rozdeliť index vnútrozhlukovej vzdialenosti na dve časti, a teda

$$J_p(\mathbf{X}, \mathbf{M}) = \sum_{h=1}^{k} \sum_{i}^{n} \left[u_{ih} \sum_{l=1}^{n_1} (x_{il} - c_{hl}^{mod})^2 + \gamma u_{ih} \sum_{l=n_1+1}^{m} \delta(x_{il}, c_{hl}^{mod}) \right]. \tag{2.19}$$

V tejto metóde postupujeme rovnako ako pri k-modes, až na krok prepočítavania nového centra, kde súradnice centra počítame priemerom pre prvých n_1 premenných, ako v k-means algoritme a zvyšné, nominálne, súradnice počítame podľa lemmy 2.5.1 ako v metóde k-modes. Túto upravenú metódu nazývame metóda k-prototypes [10].

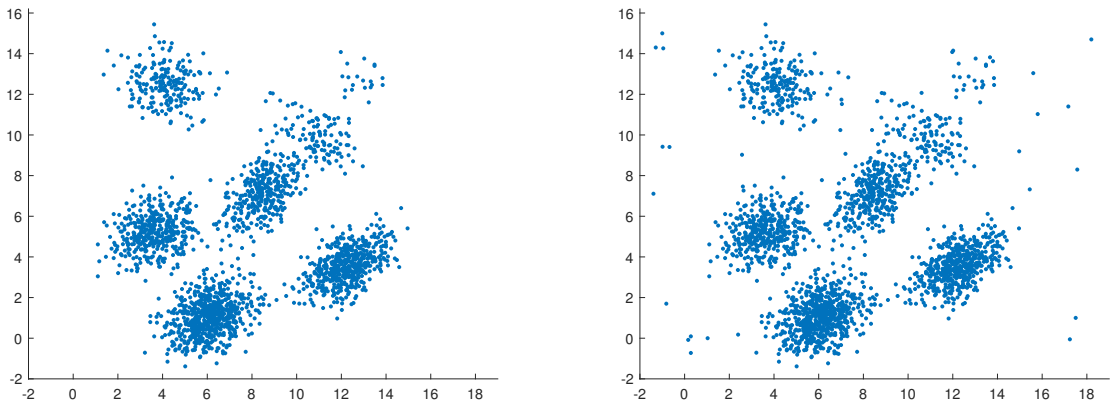
Kapitola 3

Aplikácia zhlukovej analýzy na umelé dáta

V tejto kapitole sa budeme zaoberať aplikáciou metód zhlukovej analýzy na simulované dáta, ktoré sme vytvorili za účelom ilustrácie použitia nástrojov zhlukovej analýzy. Všetky použité algoritmy sme implementovali v prostredí MATLAB. Pre zjednodušenie znázornenia sme zvolili dvojrozmerné dáta s normálnym rozdelením s hustotou pravdepodobnosti

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} exp(-\frac{(x_1-\mu_1)^2}{2\sigma_1} - \frac{(x_2-\mu_2)^2}{2\sigma_2}), \tag{3.1}$$

kde parametre σ_1, σ_2 sú vyberané náhodne (s ohľadom na použitú funkciu *rand()*). Budeme pracovať s 2260 objektmi, ktoré majú dva príznaky (súradnice x a y) ako s dátami bez šumu znázornenými na obrázku 3.1a, resp. s 2314 objektmi o dvoch príznakoch, ktoré tvoria dáta so šumom na obrázku 3.1b.



(a) Dáta bez šumu

(b) Dáta s pridaním uniformného šumu.

Obr. 3.1: Simulované dvojrozmerné dáta s normálnym rozdelením, na ktorých ilustrujeme implementované metódy.

Podľa postupu uvedenom v kapitole 1, započneme štvorkrokový proces voľbou miery podobnosti. Pre názornú ilustráciu volíme euklidovskú vzdialenosť (1.3), navyše pre porovnanie blokovú (1.2) a logaritmickú(1.4). Druhým krokom je výber vhodnej metódy. Keďže sa jedná

o ilustratívnu kapitolu, ukážeme metódy k-means, k-means++, dvojfázový k-means i metódu k-medoids. Ďalším, dôležitým krokom je určenie počtu zhlukov a nakoniec interpretácia výsledkov. Záver tejto kapitoly obohatíme o niekoľko experimentov s rôznymi tvarmi vnútrozhlukových a mimozhlukových energií.

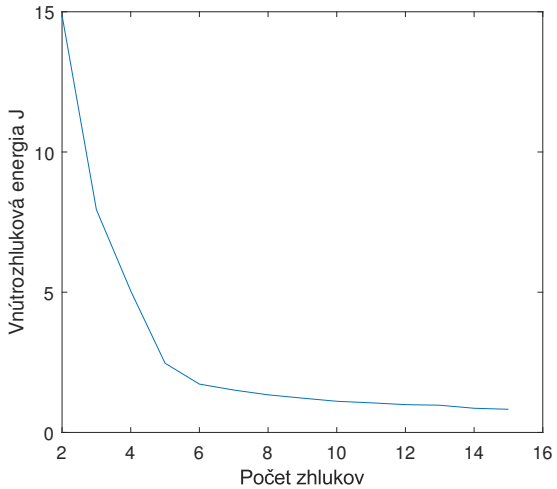
3.1 Určenie počtu zhlukov k

Počet zhlukov vrámci zhlukovania možno určiť pomocou rôznych metód, vrátane toho, že si užívateľ túto hodnotu jednoducho tipne. Rozhodli sme sa popísať určenie k dvomi spôsobmi: metódou kolena (lakťa) [8] a metódou obrysového koeficientu [7].

Metóda kolena

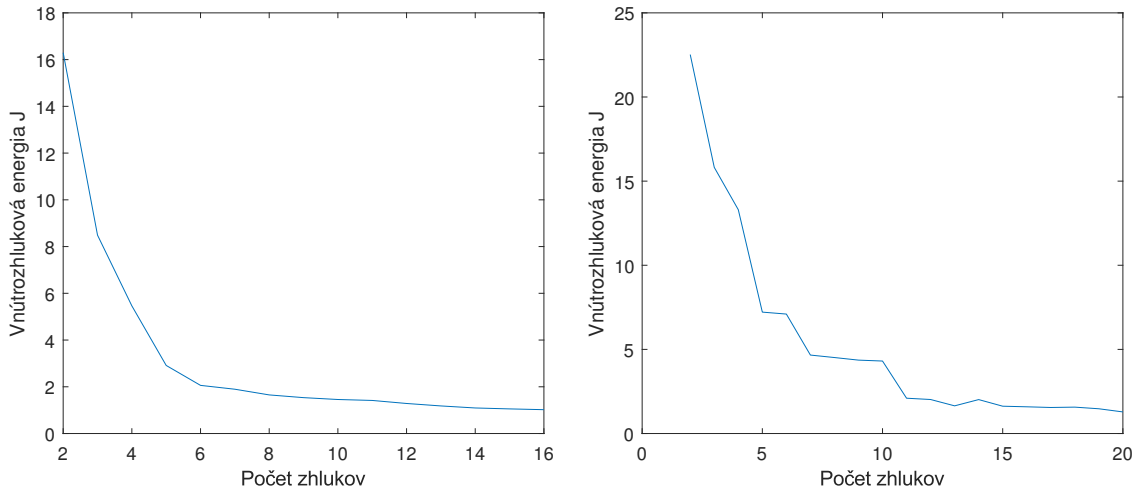
Metóda kolena (z angl. *elbow rule*) [8] sa zameriava na vnútrozhlukovú mieru resp. energiu J (najmä vo forme WCSS (2.2)). Princípom je postupne prechádzať proces zhlukovania na požadovanom datasete pre rôzne hodnoty počtu zhlukov $k = 1, 2, \ldots, \tilde{k}$, pričom nás zaujíma hodnota vnútrozhlukovej miery. V procese zhlukovania sa snažíme minimalizovať túto mieru J, preto nám nižšie hodnoty ukazujú vyššiu konvergenciu zhlukov. Poznamenajme, že sa nesnažíme dosiahnuť globálne minimum, pretože globálne minimum $J = 0$ nastáva pre $k = n$, teda ak každý objekt stojí samostatne, čo je pre nás nezaujímavým riešením.

Keď sa blížime k skutočnej hodnote počtu zhlukov k, ukazuje sa, že dôjde k rapídnemu poklesu v hodnote J. Následne ak sa vzďaľujeme od skutočnej hodnoty k (počet zhlukov narastá), vnútrozhluková miera síce klesá, avšak tento pokles je badateľne miernejší. Výhodné je vývoj energie J graficky zobraziť a nájsť takzvané *plató* na krivke energie (grafické znázornenie pripomína končatinu ohnutú v kĺbe: v lakti či kolene).



Obr. 3.2: Vývoj vnútrozhlukovej miery J (WCSS (2.2) normovanej počtom objektov) vzhľadom na počet zhlukov k pre simulovaný dataset bez šumu 3.1a v metóde k-means.

Na základe metódy kolena aplikovanej na dataset bez šumu pri použití algoritmu k-means môžeme z grafu 3.2 vnútrozhlukovej miery usúdiť, že "koleno"sa na krivke nachádza pre $k = 6$. Túto domnienku overíme v nasledujúcej sekcií metódou obrysového koeficientu 3.1.



(a) (b)

Obr. 3.3: (a)Vývoj vnútrozhlukovej miery J (WCSS (2.2) normovanej počtom objektov) vzhľadom na počet zhlukov k pre simulovaný dataset so šumom na obrázku 3.1b v metóde k-means.
(b)Vývoj vnútrozhlukovej miery J (WCSS (2.2) normovanej počtom objektov) vzhľadom na počet zhlukov k pre simulovaný dataset so šumom na obrázku 3.1b v metóde dvojfázového k-means.

Pre dataset so šumom nastane situácia, kedy podľa metódy kolena, s využitím algoritmu k-means, sme na krivke energie J 3.3a schopní určiť koleno pre $k = 6$. Pokiaľ však k určeniu počtu zhlukov v tejto metóde využijeme algoritmus dvojfázového k-means, na grafe 3.3b, nemôžeme s istotou povedať, že $k = 4$, $k = 7$ alebo $k = 11$.

Obrysový koeficient

Iným spôsobom je využitie takzvaného Obrysového koeficientu (z angl. *Silhouette Coefficient*) [7], ktorý porovnáva vnútrozhlukovú mieru s medzizlukovou. Pre každý objekt $\mathbf{x}_i$ chceme určiť šírku obrysu (z angl. *Silhouette Value*), ktorá nesie informáciu o tom, ako veľmi je objekt podobný svojmu vlastnému zhluku (kohézia, ucelenosť) v porovnaní s ostatnými nevlastnými zhlukmi (separácia). Šírku obrysu pre každý objekt $\mathbf{x}_i$ vypočítame nasledovne:

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}, \tag{3.2}$$

kde a(i) značí priemer vzdialeností medzi objektom $\mathbf{x}_i$ a ostatnými prvkami toho istého zhluku, a koeficient b(i) predstavuje minimum priemeru vzdialeností medzi objektom $\mathbf{x}_i$ a všetkými objektami z nevlastných zhlukov, inými slovami priemernú vzdialenosť od najbližšieho cudzieho zhluku. Ak $\mathbf{x}_i \in \mathcal{C}_q$ a ak označíme množinu chudzích zhlukov ako $\tilde{\mathcal{C}} = \{\mathcal{C}_h : h \in \widehat{k} \setminus \{q\}\}$, môžeme písať:

$$a(i) = \frac{1}{|\mathcal{C}_q - 1|} \sum_{\mathbf{x}_j \in \mathcal{C}_q} d(\mathbf{x}_i, \mathbf{x}_j), \tag{3.3a}$$

$$b(i) = \min_{\mathcal{C}_h \in \tilde{\mathcal{C}}} \frac{1}{|\mathcal{C}_h|} \sum_{\mathbf{x}_j \in \mathcal{C}_h} d(\mathbf{x}_i, \mathbf{x}_j). \tag{3.3b}$$

Obrysová šírka nadobúda hodnoty od -1 po 1 ($s(i) \in \langle -1; 1 \rangle$). Ak je hodnota s(i) blízka -1, objekt je nesprávne zaradený; ak je hodnota s(i) v okolí 0, znamená to, že objekt $\mathbf{x}_i$ je

možné rovnako dobre zaradiť aj k inému zhluku; a ak sa s(i) nachádza blízko druhej krajnej hodnoty 1, možno tvrdiť, že objekt je zaradený správne. Tieto skutočnosti sa využívajú pri tvorbe takzvaného obrysového grafu, ktorého výhoda spočíva v neobmedzenosti dimenzie prvkov. Tento graf bližšie popisujeme v sekcií 3.2.3, kde ukážeme jeho použitie na simulovaných dátach a budeme ho využívať v kapitole 4.

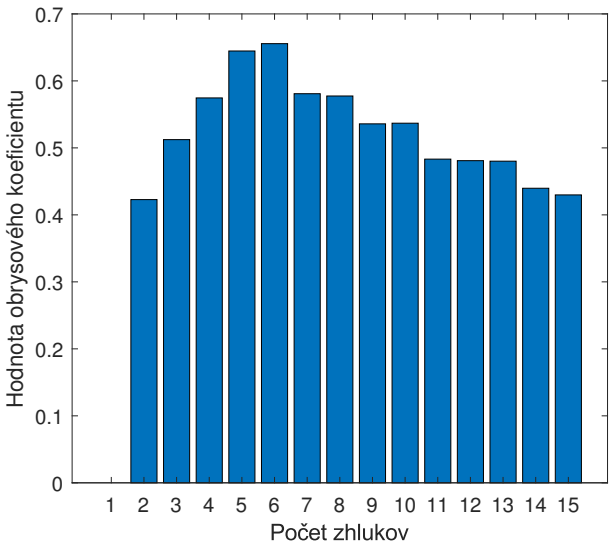
Aby sme mohli charakterizovať kvalitu zhlukovania ako celku, zostavíme obrysový koeficient. Pre konkrétny počet zhlukov k, možno spočítať obrysový koeficient ako priemernú hodnotu šírok obrysov

$$\overline{s}(k) = \frac{1}{n}\sum \mathbf{x}_i \in \mathbf{X} s(i) = \frac{1}{n}\sum_{i=1}^{n} s(i). \tag{3.4}$$

Vďaka tomuto koeficientu môžeme rozhodnúť o správnej hodnote počtu zhlukov. Aplikujeme zvolený zhlukovací algoritmus na dataset pre sériu rôznych hodnôt počtov zhlukov $k = 1, 2, \ldots, \tilde{k}$, pričom skutočný počet zhlukov určíme podľa maximálnej nadobudnutej hodnoty obrysového koeficientu. Ilustratívne možno zapísať formou

$$k = \arg\max_{k \in \widehat{\tilde{k}}} \overline{s}(k). \tag{3.5}$$

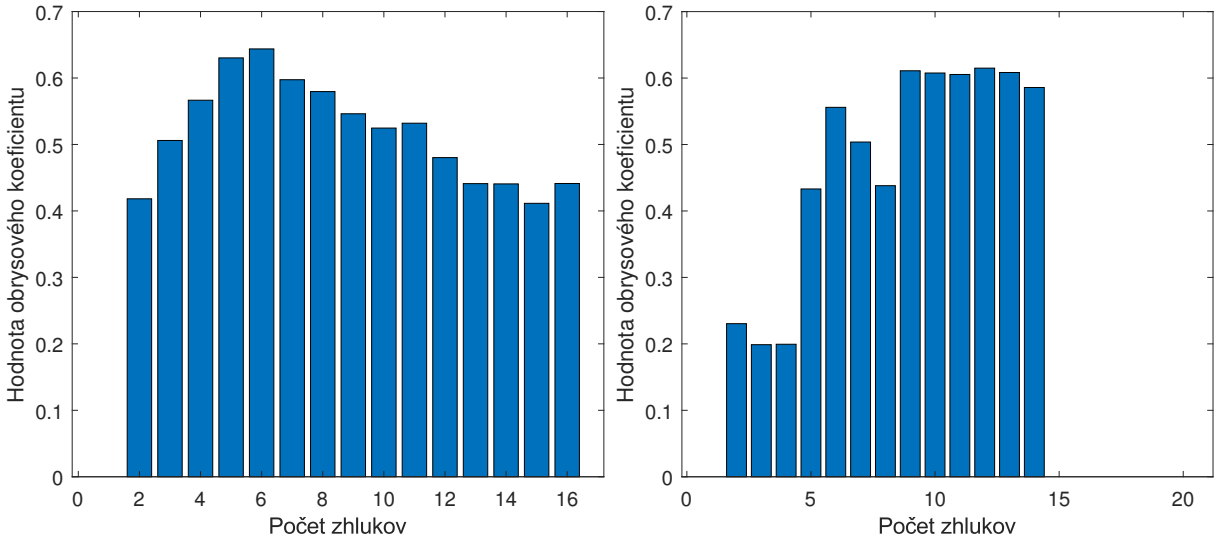
Vyššie popísanú metódu sme implementovali v prostredí MATLABu. Postupne sme pre rôzne počty zhlukov $k = 2, 3, \ldots, 15$ na dataset aplikovali metódu k-means s užitím euklidovskej metriky (1.3), tak, že sme algoritmus opakovali pätnásťkrát s náhodnými inicializáciami a do ďalšieho výpočtu sme posunuli len výsledok, kde rozdelenie do zhlukov poskytlo najnižšiu hodnotu vnútrozhlukovej miery (2.2) normovanej na jeden prvok. Následne sme pre každý objekt vypočítali jeho obrysovú šírku (3.2) a z ich priemeru sme dostali obrysový koeficient (3.4) pričom najvyššia hodnota $\overline{s}(k) = 0.6556$ zodpovedala $k = 6$.



Obr. 3.4: Graf hodnôt obrysového koeficientu vzhľadom k počtu zhlukov pri aplikácií metódy k-means na dataset bez šumu.

Zamerali sme sa aj na dataset so šumom, kedy sme najprv využili metódu k-means. Po spočítaní obrysových šírok sme hodnoty pre jednotlivé hodnoty počtu zhlukov spriemerovali a našli najvyššiu hodnotu obrysového koeficientu $\overline{s}(k) = 0.6438$, ktorá zodpovedala $k = 6$. Znázornenie jednotlivých šírok je možné vidieť na 3.5a. Túto metódu sme sa rozhodli nasadiť aj

na dvojfázový k-means. Najvyššia hodnota obrysového koeficientu, zostaveného podľa popisu, $\overline{s}(k) = 0.6194$ náleží $k = 12$.



(a) (b)

Obr. 3.5: Grafy hodnôt obrysových koeficientov vzhľadom k počtu zhlukov pri aplikácií k-means (a); dvojfázového k-means (b) na dataset so šumom 3.1b.

3.2 Aplikácia zhlukovacích metód

Keď už máme zvolenú mieru podobnosti, rovnako aj metódy a v neposlednom rade sme podľa metódy kolena a metódy obrysového koeficientu určili, že je vhodné sa pohybovať okolo $k = 6$ v datasete bez šumu a okolo $k = 6$ resp. $k = 12$ v datasete so šumom, môžeme pristúpiť k aplikácií metód. Rozhodli sme sa pre každú z metód (k-means, k-means++, dvojfázový k-means, k-medoids) opakovať algoritmus stojedenásťkrát, zároveň poznamenajme, že algoritmy začínajú náhodnou inicializáciou centier (s ohľadom na náhodnosť generátorov programu MATLAB). V prípade dát bez šumu demonštrujeme nielen zhlukovanie pre vhodne zvolené k, ale ukážeme si prípady, kedy je počet zhlukov podhodnotený, respektíve nadhodnotený. Vrámci dát so šumom sa obmedzíme na počty zhlukov stanovených metódou kolena a nadhodnotením počtu zhlukov.

Naše experimenty na dataset bez šumu prebehnú pre $k = 3$, $k = 5$, $k = 6$, $k = 7$ a $k = 9$, pri datasete so šumom volíme najprv $k = 6$ pre jednofázové metódy a $k = 12$ pre dvojfázovú metódu, následne navýšime počet zhlukov pre obe skupiny metód na $k = 13$, aby sme nahliadli do hlavného rozdielu v týchto dvoch prístupoch. Výsledky zaznamenáme do tabuliek. V stojedenástich opakovaniach si zaznamenávame výsledné hodnoty vnútrozhlukovej i mimozhlukovej miery, ku ktorým v danej inicializácii algoritmus dokonvergoval. Vnútrozhlukovú mieru J počítame podľa (2.2) a normujeme ju na jeden prvok, mimozhlukovú mieru E počítame podľa prístupu metódy priemernej väzby, ktorú sme popísali v (1.11). Do tabuľky následne zaznamenáme minimálnu hodnotu J a maximálnu hodnotu E, priemernú hodnotu mier E a J, a taktiež hodnotu ich mediánov. Na základe týchto veličín môžeme vykresliť takzvaný boxplot, ktorým vizualizujeme hodnoty energií okolo mediánu a prípadné odľahlé merania (taká inicializácia, ktorá mala príliš rozdielnu energiu). V rámci použitia euklidovskej metriky vykresľujeme finálne zaradenia algoritmov pre iteráciu, ktorá dosiahla minimálnu hodnotu J a iteráciu, ktorá dosiahla maximálnu hodnotu E. Vykreslenie riešenia pre maximálne dosiahnuté E je čisto pre zaujíma-

vosť, pretože implementované algoritmy sú založené na minimalizácií WCSS (2.2). Vizualizáciu riešenia s použitím blokovej a logaritmickej vzdialenosti možno nájsť v prílohe 4.2.3.

3.2.1 Dataset bez šumu

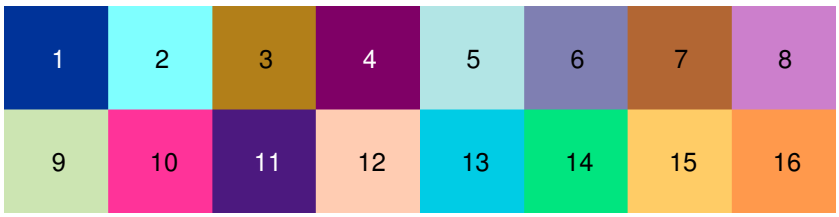
k=3

V prvej demonštrácií sa zameriame na podhodnotený počet zhlukov, konkrétne $k = 3$. Z tabuľky 3.1 vidíme, že riešenia algoritmov k-means a k-means++ energeticky dokonvergovali k rovnakej hodnote $J_{min} = 7,94$ v rámci použitia euklidovskej metriky. Priemerne mali najnižšie hodnoty energií výsledky algoritmu k-means++ a to bez ohľadu na použitú metriku.

Naopak dvojfázový k-means, dosahoval priemerne najneuspokojivejšie hodnoty J, v porovnaní s ostatnými použitými metódami. Nevýhodou tohto algoritmu bol okrem iného vyšší výpočetný čas a fakt, že implementácia algoritmu (konkrétne fáza 1) podľa nášeho popisu 7 často skĺzne k zacykleniu. Z tohto dôvodu sme stanovili hraničný počet iterácií cyklu prvej fázy na 200 iterácií (s ohľadom na odpozorované správanie ostatných implementácií, kedy sa algoritmy k-means, k-means++, k-medoids ustália do 40 iterácií). Tento nežiaduci efekt je spojený so zmenou počtu zhlukov vo fáze 1. Rozhodli sme sa, že z dôvodu absencie šumu, na pohľad odľahlých prvkov a značnému výpočetnému času tento algoritmus nebudeme demonštrovať pre ďalšie hodnoty k, okrem $k = 7$.

Ďalej si môžeme všimnúť, že ani jedno z výsledných rozdelení nie je na pohľad uspokojivé, zaradenie hraničných prvkov považujeme prinajmenšom za sporné. V dôsledku podhodnotenia počtu zhlukov sa centrá vzniknutých zhlukov tvoria miestach, kde neležia žiadne objekty (medzi opticky viditeľnými zhlukmi). Preto pozorujeme, že objekty na periférii niektorých zhlukov by mali spadať k zhlukom iným. Hranica zhlukov vo výsledku nerešpektuje vizuálne rozloženie, ktoré ľudským okom vidíme.

Na obrázku 3.6 mapujeme farby zhlukov k číslam zhlukov, pričom číslo zhluku je udané uprostred farebného štvorca, aby sme mohli prepojiť farebné vyobrazenie riešení zhlukovania z častí sekcie 3.2 so zhlukmi udanými na obrysových grafoch v časti 3.2.3. Táto paleta je konštantná a v obrysovom grafe sú zhluky v poradí od 1 do k.



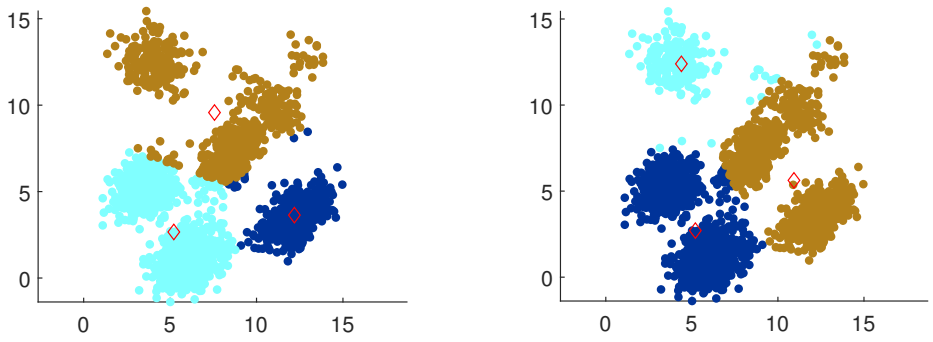
Obr. 3.6: Legenda zafarbenia jednotlivých zhlukov popísaná číslom zhluku.

	euklidovská metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	7,94	7,94	8,11	8,93
J_{priem}	8,24	8,18	9,84	9,38
J_{med}	7,95	8,10	9,26	9,17
E_{max}	54,10	54,10	54,30	46,96
E_{priem}	47,46	49,71	47,78	39,27
E_{med}	47,08	47,09	45,58	37,39

	bloková metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	13,05	13,17	13,23	14,13
J_{priem}	13,58	13,53	15,27	14,75
J_{med}	13,69	13,69	13,24	14,78
E_{max}	68,65	68,65	68,73	66,50
E_{priem}	60,38	61,69	67,82	60,72
E_{med}	60,09	60,09	68,63	59,82

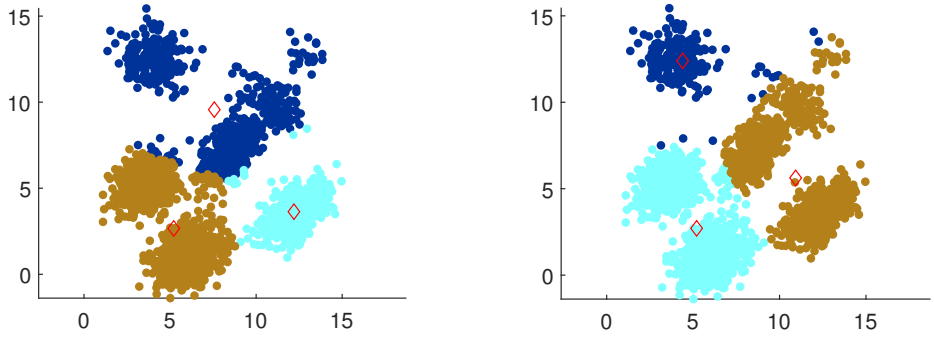
	logaritmická metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	4,14	4,14	4,40	4,15
J_{priem}	4,34	4,30	5,00	4,40
J_{med}	4,40	4,29	4,40	4,33
E_{max}	25,88	25,88	25,91	25,27
E_{priem}	24,35	24,34	25,75	24,08
E_{med}	24,15	24,15	25,91	23,90

Tabuľka 3.1: Tabuľka algoritmov k-means, k-means++, dvojfázového k-means, k-medoids pre $k = 3$, ktoré prebehli 111 opakovaní s náhodnými inicializáciami. Obsahuje najnižšie nadobudnuté hodnoty vnútrozhlukovej miery J, maximálne nadobudnuté hodnoty medzizhlukovej miery E, priemerne dosiahnuté hodnoty vnútrozhlukovej J a medzizhlukovej E miery a ich mediány pre rôzne druhy vzdialeností.



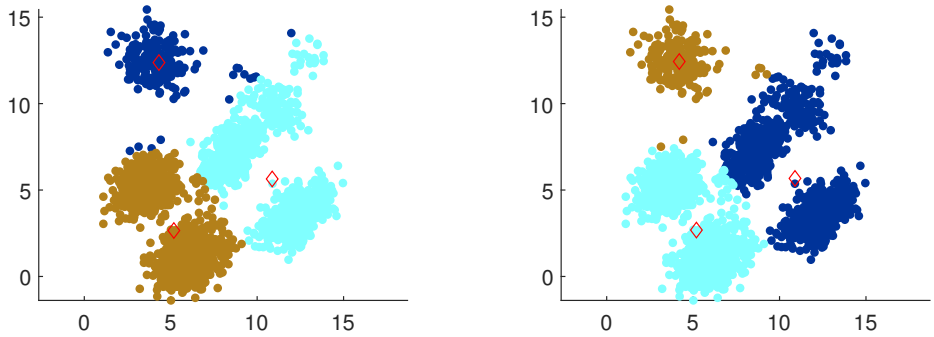
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.7: Výsledky zhlukovacieho procesu k-means, $k = 3$ pre energie uvedené v tab. 3.1.



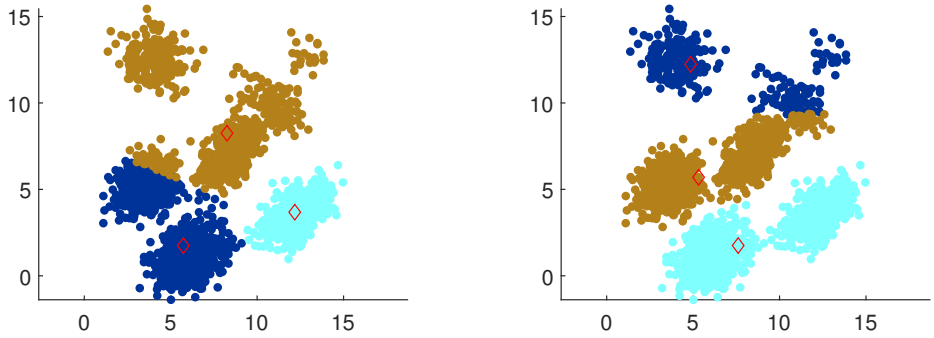
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.8: Výsledky zhlukovacieho procesu k-means++, $k = 3$ pre energie uvedené v tab. 3.1.



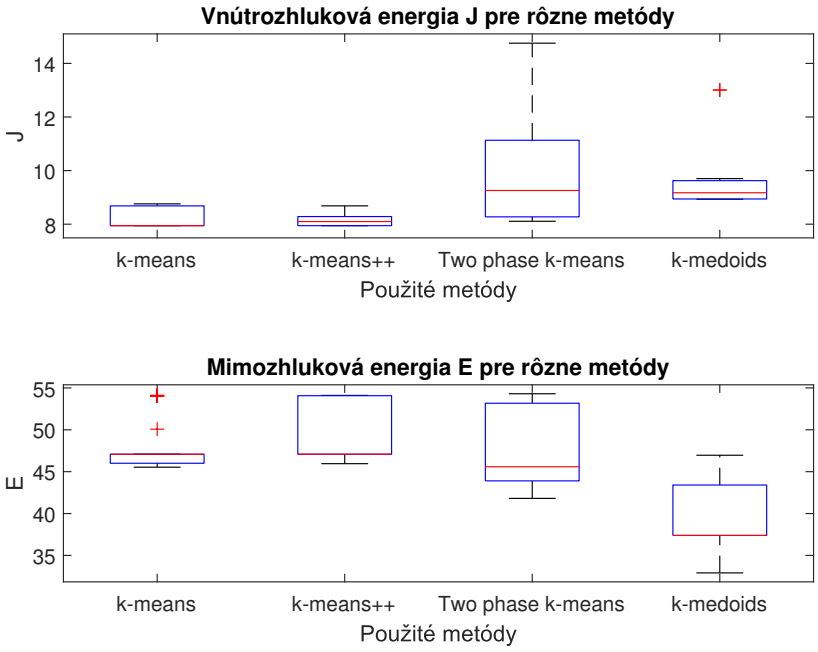
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.9: Výsledky zhlukovacieho procesu dvojfázového k-means, $k = 3$ pre energie uvedené v tab. 3.1.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.10: Výsledky zhlukovacieho procesu k-medoids, $k = 3$ pre energie uvedené v tab. 3.1.



Obr. 3.11: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 3$ v priebehu 111 behov jednotlivých algoritmov s náhodnými inicializáciami.

k=5

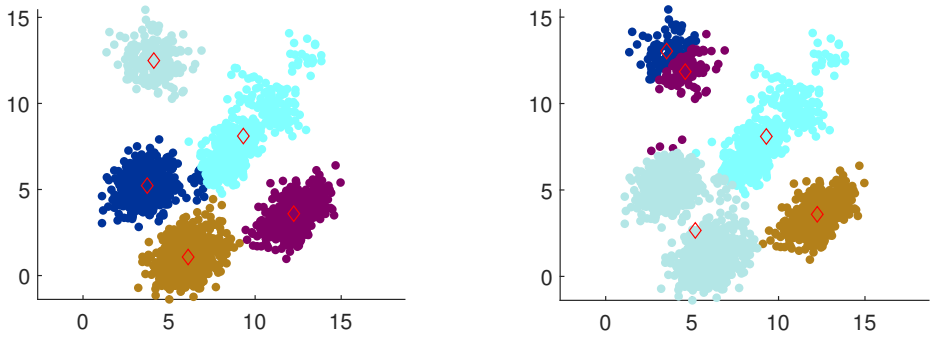
V ďalšej ukážke sa zameriame na mierne podhodnotený počet zhlukov oproti teoretickému odhadu a v dátach budeme hľadať 5 zhlukov. Tentokrát môžeme z tabuľky 3.2 vyčítať, že k-means a k-means++ dokonvergovali v euklidovskej metrike k rovnakému minimálnemu riešeniu s $J_{min} = 2,47$ (rovnako tomu je aj pre ostatné metriky) a algoritmus k-medoids sa k tomuto riešeniu značne priblížil a dosiahol $J_{min} = 2,56$. Môžeme si všimnúť, že vrámci použitie logaritmickej miery je rozdiel medzi minimálne dosiahnutou energiou J pre k-means resp. k-means++ a k-medoids len jedna stotina (pre euklidovskú a blokovú metriku desatina).
Ak porovnáme vizualizácie k-means a k-medoids pre minimálnu hodnoru J, vidíme, že pozdĺžny zhluk (tyrkysovej farby) má pre k-means 3.12a mierne vychýlené centrum smerom k pravému hornému rohu, pričom v metóde k-medoids má tento pozdĺžny zhluk 3.14a ukotvenie nižšie. Ako sme spomínali v sekcií 2.3, medoidy nie sú toľko náchylné k vychýleným objektom, preto sa oproti k-means algoritmu centrum neposúva nahor k odľahlejšiemu podzhluku a ukotvil sa v časti s vyšším počtom objektov. Táto poloha následne ovplyvňuje hodnoty vnútrozhlukový mier.

	euklidovská metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	2,47	2,47	2,56
J_{priem}	3,26	2,78	4,37
J_{med}	2,47	2,47	5,40
E_{max}	162,81	158,05	165,72
E_{priem}	154,08	156,58	148,88
E_{med}	158,05	158,05	155,41

	bloková metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	4,21	4,21	4,35
J_{priem}	5,57	4,76	7,08
J_{med}	4,20	4,21	7,93
E_{max}	201,07	204,18	211,17
E_{priem}	194,49	198,66	189,07
E_{med}	201,06	201,06	192,93

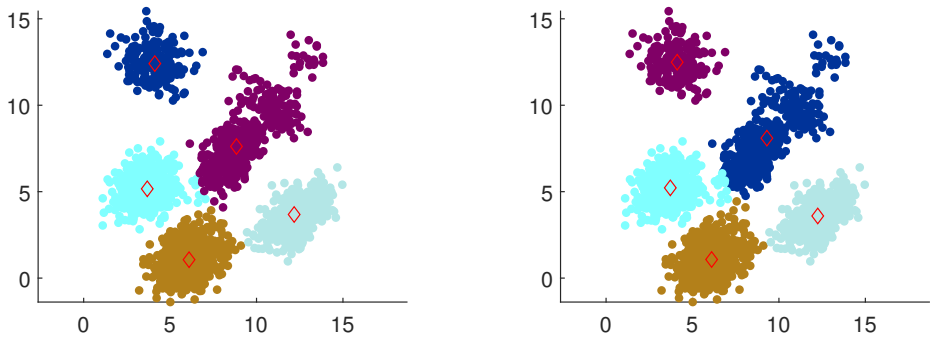
	logaritmická metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	1,37	1,37	1,38
J_{priem}	1,86	1,53	2,20
J_{med}	1,38	1,38	2,51
E_{max}	80,72	80,72	80,72
E_{priem}	78,61	80,19	77,57
E_{med}	80,71	80,71	78,12

Tabuľka 3.2: Tabuľka algoritmov k-means, k-means++, k-medoids pre $k = 5$, ktoré prebehli 111 opakovaní s náhodnými inicializáciami. Obsahuje najnižšie nadobudnuté hodnoty vnútrozhlukovej miery J, maximálne nadobudnuté hodnoty medzizhlukovej miery E, priemerne dosiahnuté hodnoty vnútrozhlukovej J a medzizhlukovej E miery a ich mediány pre rôzne druhy vzdialeností.



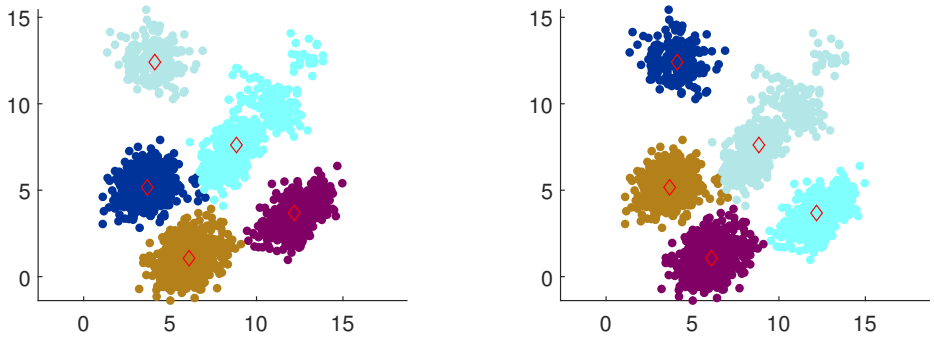
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.12: Výsledky zhlukovacieho procesu k-means, $k = 5$ pre energie uvedené v tab. 3.2.



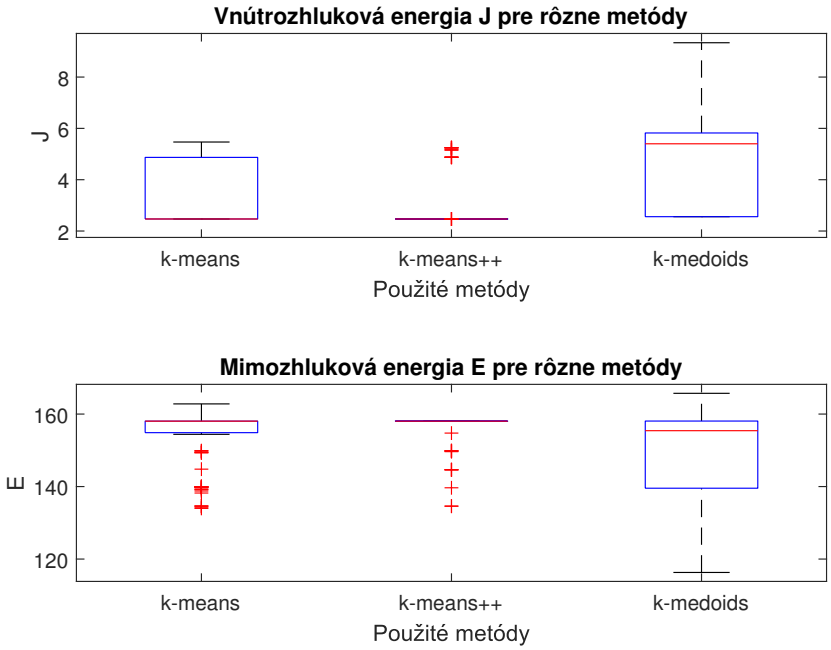
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.13: Výsledky zhlukovacieho procesu k-means++, $k = 5$ pre energie uvedené v tab. 3.2.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.14: Výsledky zhlukovacieho procesu k-medoids, $k = 5$ pre energie uvedené v tab. 3.2.



Obr. 3.15: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 5$ v priebehu 111 behov jednotlivých algoritmov s náhodnými inicializáciami.

k=6

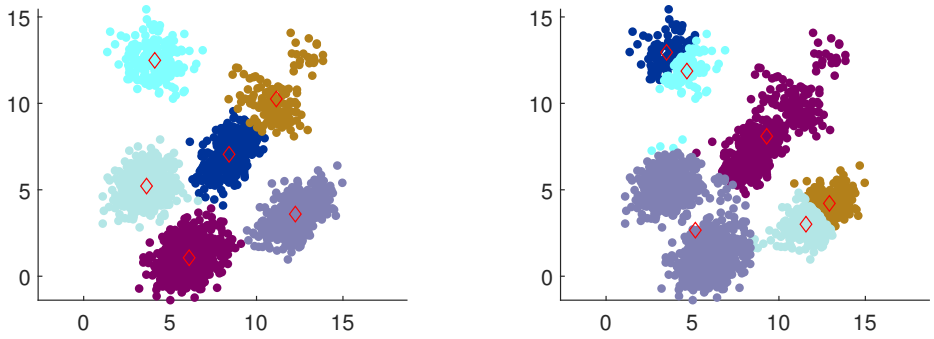
Teraz si ukážeme výsledky zhlukovania pre hodnotu počtu zhlukov stanovenú v časti 3.1 na $k = 6$. Podobne ako v predošlom prípade, metódy dosahujú podobné výsledné energie. Opakujúci trend nám ukazuje, že priemerne metóda k-medoids poskytuje vyššie hodnoty vnútrozhlukovej energie (rozdiel v stotinách) a rovnako aj vyššie hodnoty mimozhlukovej energie. Môžeme si všimnúť, že na obrázkoch 3.16, 3.17 a 3.18 všetky metódy považujú pravý horný dvojzhluk za jeden, čo je ovplyvnené práve nízkym počtom objektov v najmenšom podzhluku, tým pádom jeho príspevok k vnútrozhlukovej energii je nízky, rovnako ako príspevok jeho obrysových šírok.

	euklidovská metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	1,72	1,72	1,73
J_{priem}	2,41	2,00	3,44
J_{med}	2,25	1,72	2,39
E_{max}	241,67	240,24	242,09
E_{priem}	225,80	227,93	216,99
E_{med}	228,61	228,73	225,26

	bloková metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	2,88	2,88	2,90
J_{priem}	3,82	3,25	5,49
J_{med}	3,38	2,88	4,07
E_{max}	304,56	305,20	305,94
E_{priem}	288,17	292,10	277,34
E_{med}	293,03	293,67	284,64

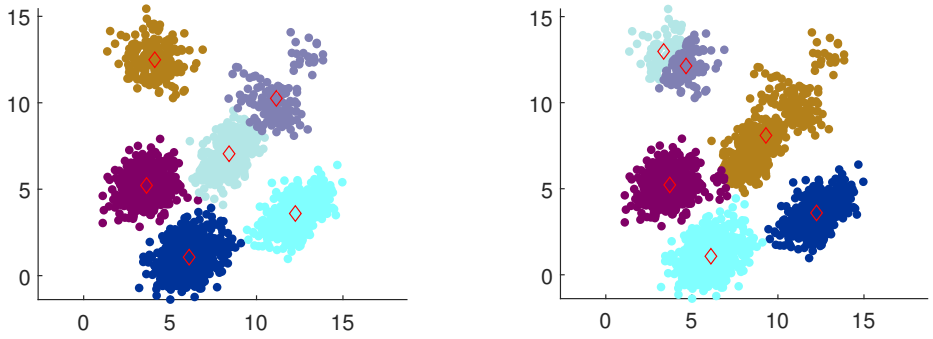
	logaritmická metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	1,00	1,00	1,00
J_{priem}	1,27	1,15	1,82
J_{med}	1,24	1,23	1,35
E_{max}	119,18	119,66	122,02
E_{priem}	116,48	117,38	112,82
E_{med}	116,91	118,61	115,95

Tabuľka 3.3: Tabuľka algoritmov k-means, k-means++, k-medoids pre $k = 6$, ktoré prebehli 111 opakovaní s náhodnými inicializáciami. Obsahuje najnižšie nadobudnuté hodnoty vnútrozhlukovej miery J, maximálne nadobudnuté hodnoty medzizhlukovej miery E, priemerne dosiahnuté hodnoty vnútrozhlukovej J a medzizhlukovej E miery a ich mediány pre rôzne druhy vzdialeností.



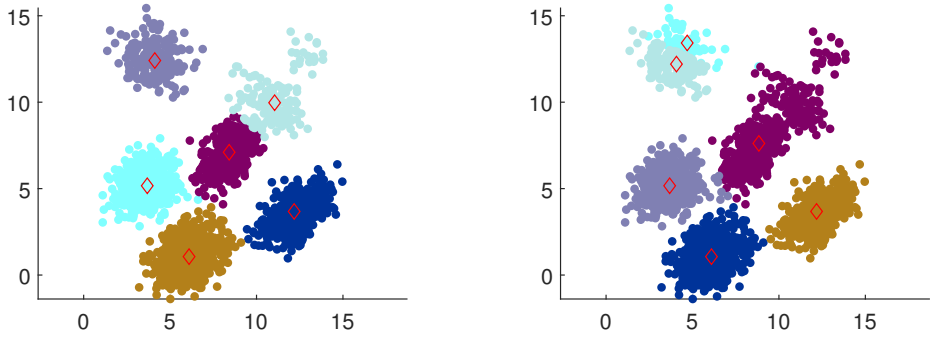
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.16: Výsledky zhlukovacieho procesu k-means, $k = 6$ pre energie uvedené v tab. 3.3.



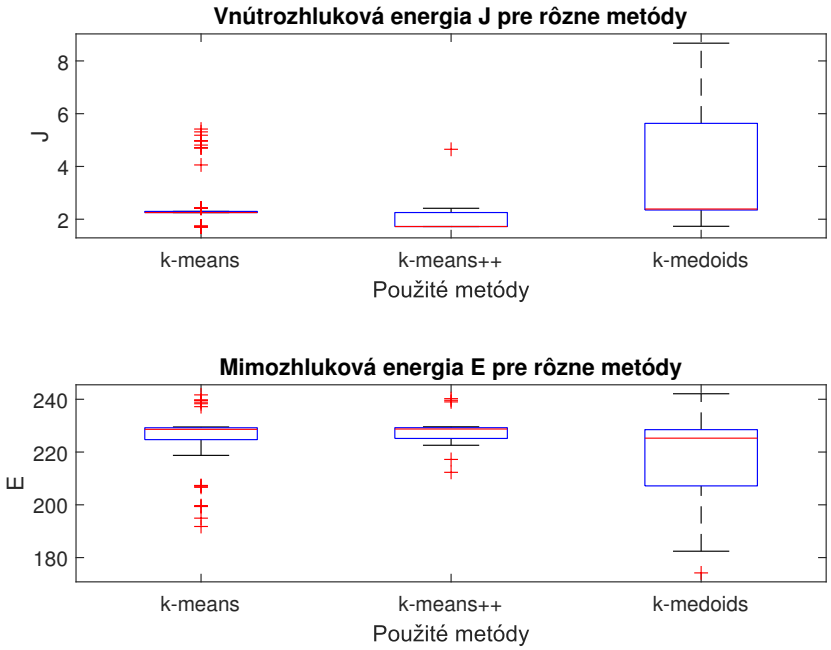
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.17: Výsledky zhlukovacieho procesu k-means++, $k = 6$ pre energie uvedené v tab. 3.3.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.18: Výsledky zhlukovacieho procesu k-medoids, $k = 6$ pre energie uvedené v tab. 3.3.



Obr. 3.19: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 6$ v priebehu 111 behov jednotlivých algoritmov s náhodnými inicializáciami.

k=7

V predošlej demonštrácií šiestich zhlukov sme si na základe grafickej interpretácie všimli existenciu dvojzhluku, preto počet zhlukov nadhodnotíme o 1. V toto prípade si môžeme všimnúť, že algoritmy k-means, k-means++ a k-medoids dosiahli hodnotu minimálnej vnútrozhlukovej energie približne 1,5, zatiaľ čo algoritmus dvojfázového k-means dosiahol hodnotu J až okolo 1,7. Pokiaľ sa zameriavame čisto na energie, môžeme tvrdiť, že riešenia prvých troch spomínaných metód sú lepšie. Ak však zhodnotíme výsledky podľa vizualizácie, všimneme si, že dvojfázový k-means oddelil práve malý horný zhluk na obrázku 3.22a a napríklad k-means++ má pre svoju minimálnu energiu J iné riešenie 3.21a. Z tohto pozorovania teda môžeme usúdiť, že napriek rôznym nástrojom je zhluková analýza v istom zmysle nejednoznačná - napriek tomu, že algoritmy minimalizujú rovnakú účelovú funkciu J, v tvare udanom WCSS (2.2), môžu algoritmy dospie k odlišným riešeniam.

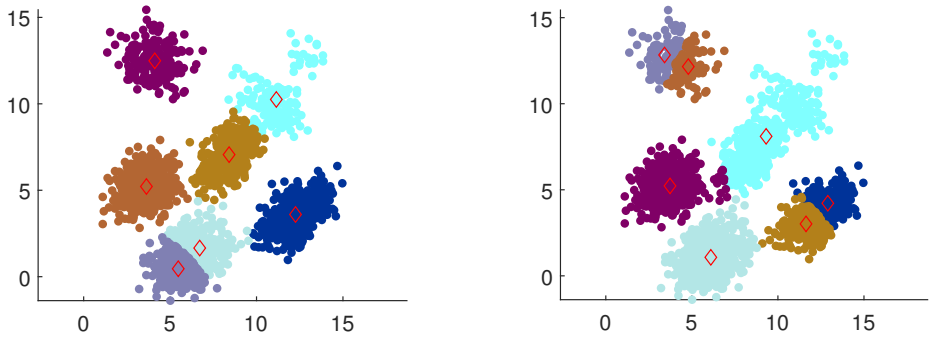
Na boxplote energií 3.24 si všimneme, že vnútrozhluková energia J pre metódu dvojfázového algoritmu k-means má v porovnaní s energiami J ostatných metód väčší rozptyl hodnôt. A naopak pre mimozhlukovú energiu E je jej rozptyl v porovnaní s ostatnými nižší. Je to spôsobené tým, že v sto jedenástich iteráciách dvojfázový k-means na základe druhej (aglomeratívnej) fázy častejšie oddelí malý horný zhluk (na obrázku 3.22a zhluk okrovej farby), pričom ostatné metódy rozdeľujú mnohopočetné zhluky. Tým pádom je napríklad príspevok energie veľkého tyrkysového zhluku v oblasti $[3, 9] \times [-1, 4]$ riešenia dvojfázového k-means (vyobrazeného na 3.22a) vyšší než súčet príspevkov belasého a fialového zhluku v tej istej oblasti riešenia k-means++ (na 3.21a).

	euklidovská metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	1,51	1,51	1,70	1,53
J_{priem}	1,80	1,65	2,93	2,78
J_{med}	1,60	1,60	2,79	2,27
E_{max}	334,21	334,82	338,48	333,64
E_{priem}	311,19	311,98	329,35	297,38
E_{med}	312,90	313,09	329,19	302,44

	bloková metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	2,50	2,51	2,69	2,54
J_{priem}	2,93	2,73	2,72	4,37
J_{med}	2,57	2,57	2,70	3,82
E_{max}	435,13	430,44	428,79	432,09
E_{priem}	397,36	401,60	423,86	385,84
E_{med}	401,83	401,86	426,60	388,71

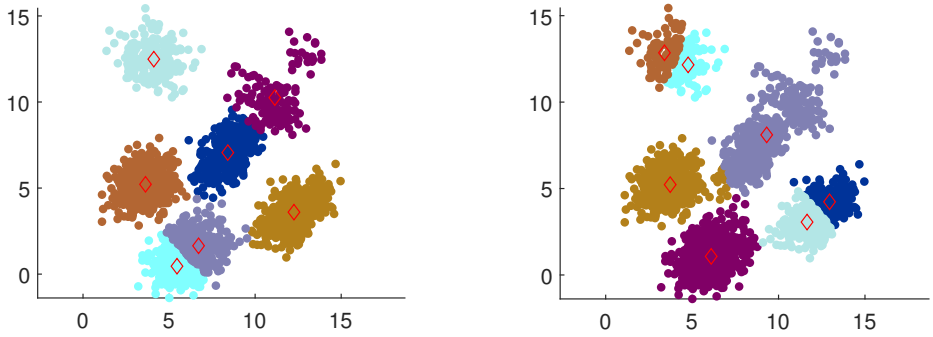
	logaritmická metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	0,86	0,86	0,94	0,87
J_{priem}	1,06	0,95	0,94	1,56
J_{med}	0,92	0,89	0,94	1,22
E_{max}	168,64	168,64	168,87	166,64
E_{priem}	159,72	161,75	168,72	154,50
E_{med}	162,39	162,51	168,73	155,72

Tabuľka 3.4: Tabuľka algoritmov k-means, k-means++, dvojfázového k-means, k-medoids pre $k = 7$, ktoré prebehli 111 opakovaní s náhodnými inicializáciami. Obsahuje najnižšie nadobudnuté hodnoty vnútrozhlukovej miery J, maximálne nadobudnuté hodnoty medzizhlukovej miery E, priemerne dosiahnuté hodnoty vnútrozhlukovej J a medzizhlukovej E miery a ich mediány pre rôzne druhy vzdialeností.



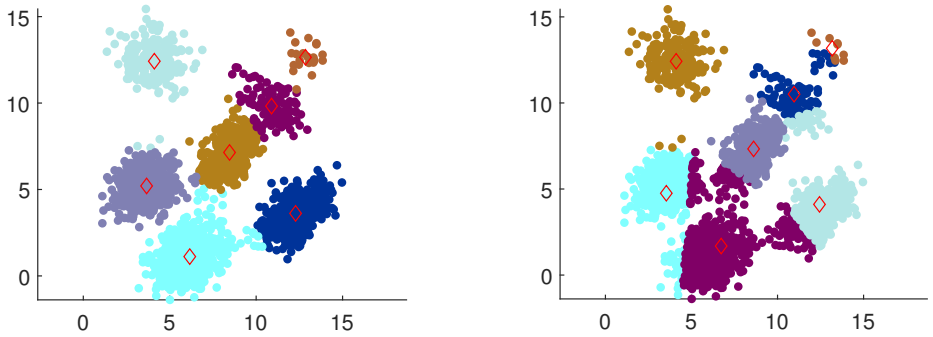
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.20: Výsledky zhlukovacieho procesu k-means, $k = 7$ pre energie uvedené v tab. 3.4.



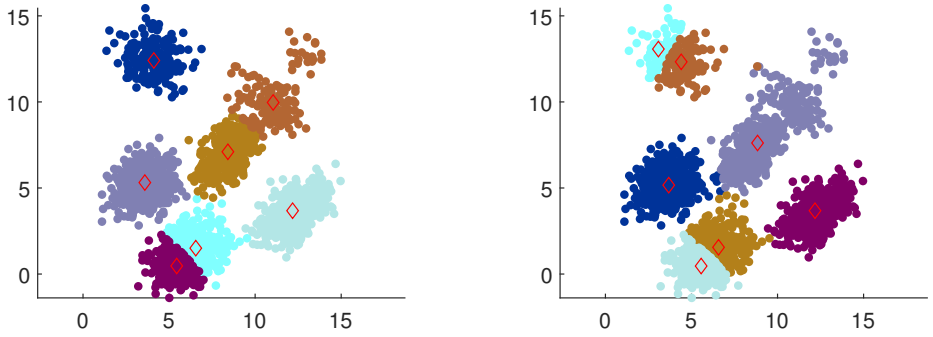
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.21: Výsledky zhlukovacieho procesu k-means++, $k = 7$ pre energie uvedené v tab. 3.4.



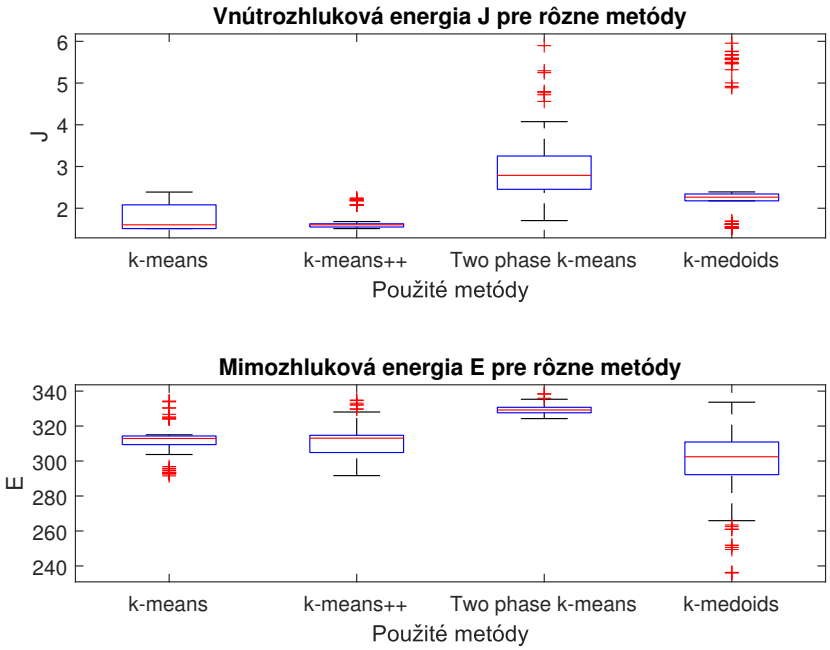
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.22: Výsledky zhlukovacieho procesu dvojfázového k-means, $k = 7$ pre energie uvedené v tab. 3.4.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

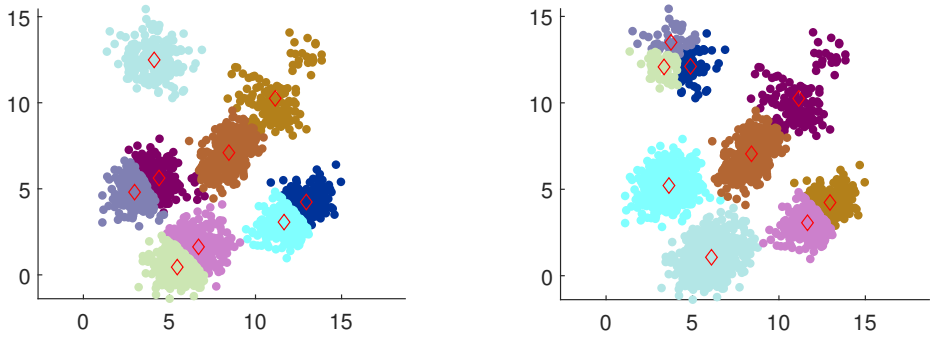
Obr. 3.23: Výsledky zhlukovacieho procesu k-medoids, $k = 7$ pre energie uvedené v tab. 3.4.



Obr. 3.24: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 7$ v priebehu 111 behov jednotlivých algoritmov s náhodnými inicializáciami.

k=9

Poslednou ukážkou na datasete bez šumu je práca s nadhodnoteným počtom zhlukov. Zvolili sme $k = 9$. V tejto časti si môžeme všimnúť, že riešenie s minimálnou vnútrozhlukovou energiou nie je na pohľad úplne prirodzené. Algoritmy sa snažia rozdeliť mnohopočetné zhluky a tým redukovať hodnotu vnútrozhlukovej energie J. Ak sa však zameriame na riešenie s maximálnou dosiahnutou hodnotou mimozhlukovej miery E, pozorujeme oddelenie malého horného zhluku napríklad na 3.26b či 3.27b. Poznamenajme však, že nami implementované algoritmy závisia svojou konvergenciou na vnútrozhlukovej miere J a mimozhlukovú mieru E uvádzame ako ďalší možný aspekt.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

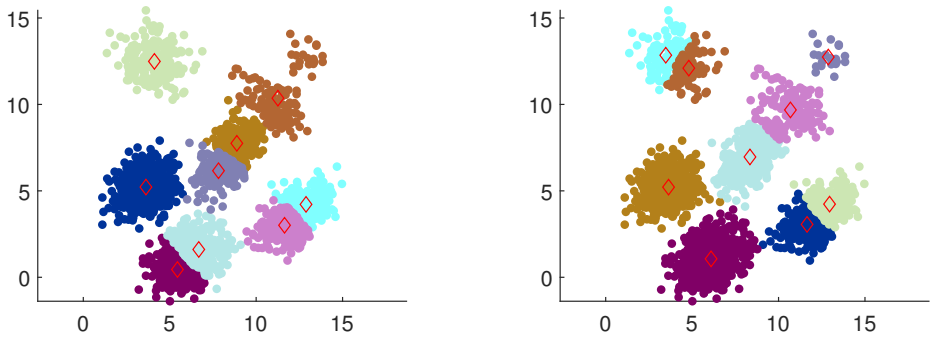
Obr. 3.25: Výsledky zhlukovacieho procesu k-means, $k = 9$ pre energie uvedené v tab. 3.5.

	euklidovská metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	1,22	1,22	1,24
J_{priem}	1,36	1,31	1,90
J_{med}	1,33	1,31	1,48
E_{max}	560,61	581,86	593,13
E_{priem}	511,08	524,37	491,65
E_{med}	511,93	517,00	482,53

	bloková metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	2,01	2,01	2,05
J_{priem}	2,16	2,16	3,18
J_{med}	2,13	2,15	2,40
E_{max}	719,54	729,27	728,48
E_{priem}	655,05	669,64	633,60
E_{med}	654,83	663,96	633,24

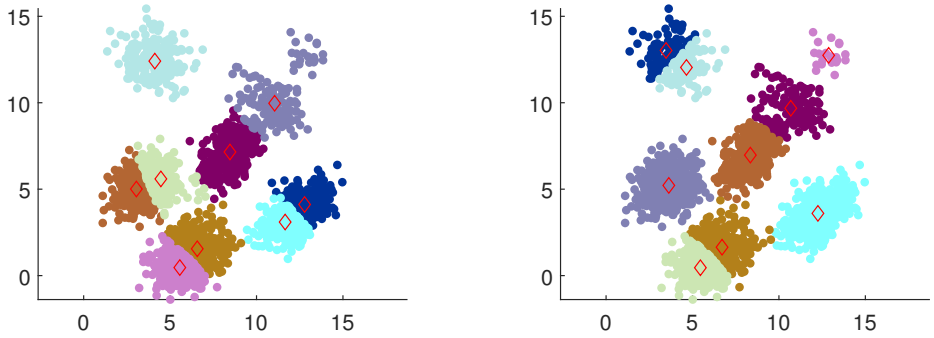
	logaritmická metrika		
Metóda	k-means	k-means++	k-medoids
J_{min}	0,67	0,67	0,68
J_{priem}	0,80	0,74	1,05
J_{med}	0,74	0,73	0,88
E_{max}	282,32	282,91	280,67
E_{priem}	266,38	270,97	258,76
E_{med}	267,13	270,66	260,69

Tabuľka 3.5: Tabuľka algoritmov k-means, k-means++, k-medoids pre $k = 9$, ktoré prebehli 111 opakovaní s náhodnými inicializáciami. Obsahuje najnižšie nadobudnuté hodnoty vnútrozhlukovej miery J, maximálne nadobudnuté hodnoty medzizhlukovej miery E, priemerne dosiahnuté hodnoty vnútrozhlukovej J a medzizhlukovej E miery a ich mediány pre rôzne druhy vzdialeností.



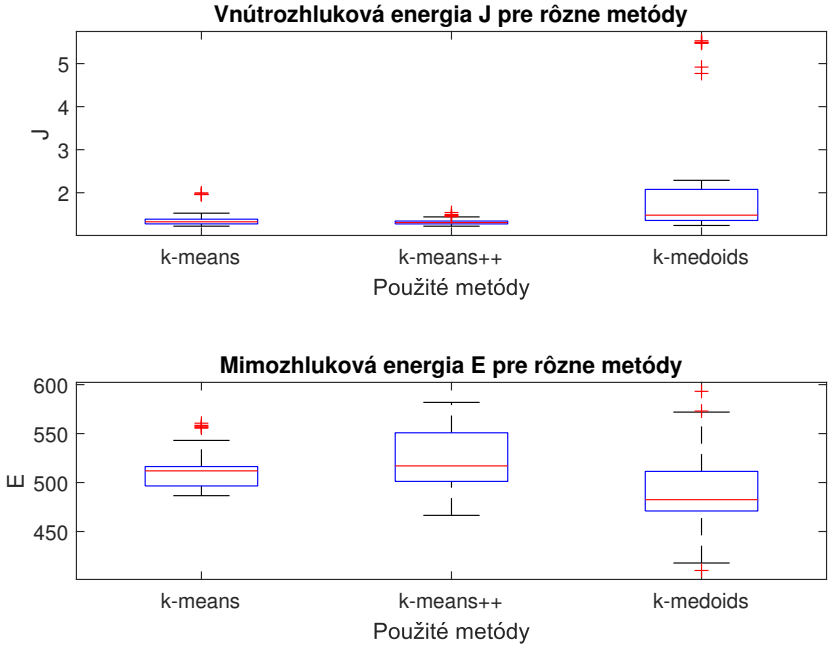
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.26: Výsledky zhlukovacieho procesu k-means++, $k = 9$ pre energie uvedené v tab. 3.5.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.27: Výsledky zhlukovacieho procesu k-medoids, $k = 9$ pre energie uvedené v tab. 3.5.



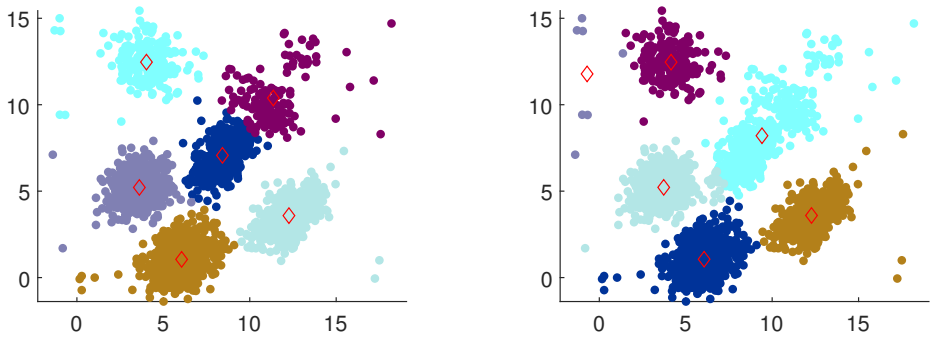
Obr. 3.28: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 9$ v priebehu 111 behov jednotlivých algoritmov s náhodnými inicializáciami.

3.2.2 Dataset so šumom

Na dataset so šumom 3.1b sme sa rozhodli v euklidovskej metrike aplikovať metódy k-means, k-means++ a k-medoids, pre ktoré sme stanovili počet zhlukov $k = 6$ na základe obrysového koeficientu 3.5a a metódu dvojfázového k-means pre $k = 12$ taktiež podľa obrysového koeficientu. Na vyobrazení riešenia k-means++, pre $k = 6$ je vidieť, že vďaka vysokej koncentrácii prvkov okolo centier, ktoré algoritmy našli, odľahlé prvky pôsobia prehliadnuté (pokiaľ by šlo napríklad o egyptologické dáta, mohlo by sa stať, že algoritmus zaradí kráľa medzi roľníkov). Pri použití dvojfázového k-means pre $k = 12$ sme očakávali detekciu odľahlých prvkov, pričom vizuálna kontrola na obrázku 3.30 ukazuje, že až na 4 ružové body odtrhnuté od svetlo-fialového zhluku vpravo hore a jedného tmavého bodu oddeleného vľavo hore prebehla detekcia úspešne. Ďalej je možné tieto odľahlé prvky buď odstrániť alebo ich ponechať ako nízkopočetné zhluky.

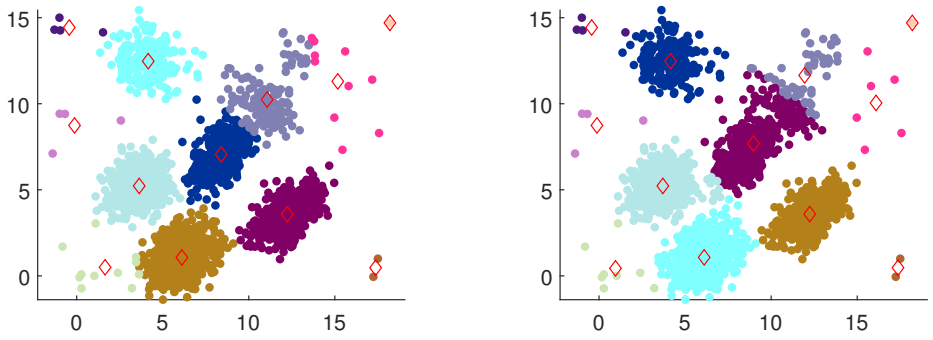
	euklidovská metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	2,06	2,06	1,77	2,08
J_{priem}	2,66	2,41	2,02	3,83
J_{med}	2,68	2,06	2,05	2,82
E_{max}	248,38	266,59	1510,40	243,60
E_{priem}	229,86	233,80	1458,79	218,40
E_{med}	232,71	232,86	1455,48	227,71

Tabuľka 3.6: Tabuľka algoritmov k-means, k-means++, k-medoids pre $k = 6$ a pre dvojfázový k-means, pre ktorý sme určili náležitý počet zhlukov $k = 12$, ktoré prebehli 111 opakovaní s náhodnými inicializáciami. Obsahuje najnižšie nadobudnuté hodnoty vnútrozhlukovej miery J, maximálne nadobudnuté hodnoty medzizhlukovej miery E, priemerne dosiahnuté hodnoty vnútrozhlukovej J a medzizhlukovej E miery a ich mediány s použitím euklidovskej vzdialenosti.



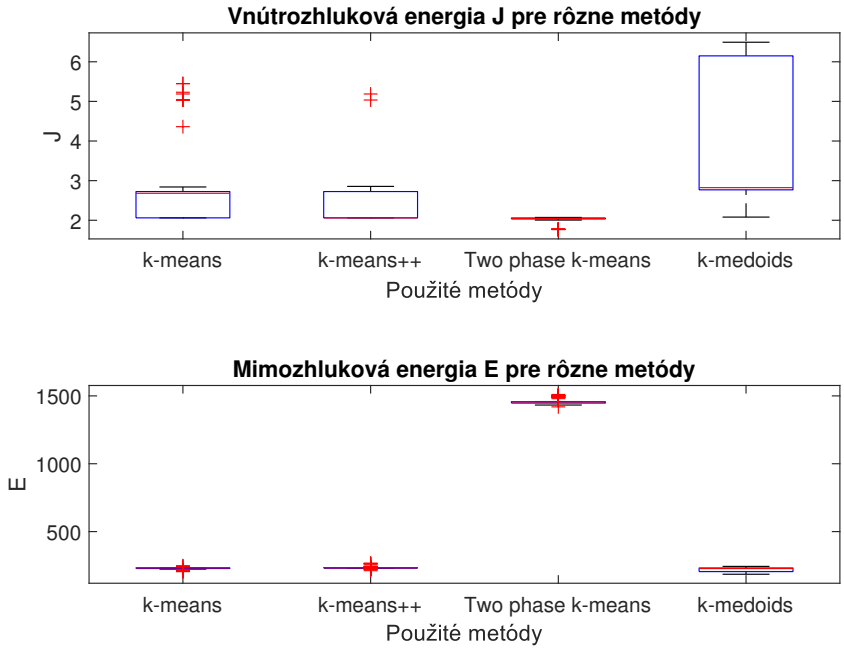
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.29: Výsledok zhlukovacieho procesu k-means++, $k = 6$ pre dáta so šumom podľa tabuľky 3.6.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.30: Výsledok zhlukovacieho procesu dvojfázového k-means, $k = 12$ pre dáta so šumom podľa tabuľky 3.6.

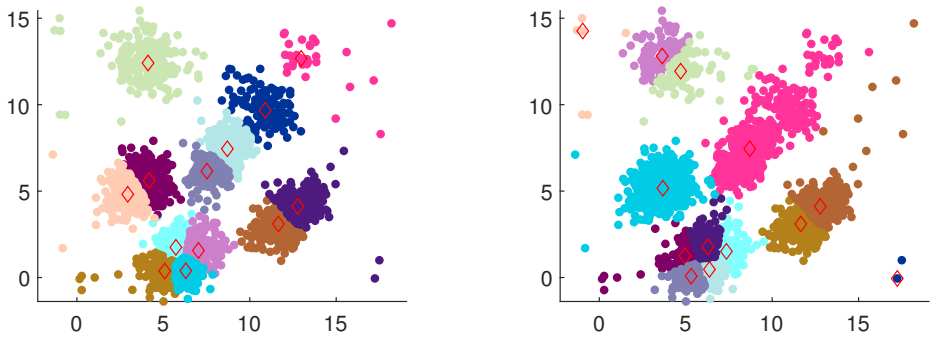


Obr. 3.31: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód k-means, k-means++, k-medoids pre $k = 6$ a dvojfázového k-means pre $k = 12$ v priebehu 111 behov jednotlivých algoritmov aplikovaných na dáta so šumom s náhodnými inicializáciami.Pri výpočtoch sa využíva euklidovská vzdialenosť.

Vrámci použitia logaritmickej vzdialenosti sme sa rozhodli navýšiť počet zhlukov na $k = 13$ pre všetky metódy. Vidíme, že dvojfázový k-means pre $k = 13$ po navýšení zhlukov o 1 oddelil problematický horný zhluk 3.33a, avšak pri rapídnom navýšení ostatných metód sa nám podobného efektu nedostalo. Na znázornení 3.32 možno postrehnúť delenie mnohopočetných zhlukov miesto separácie samostatných prvkov. Zdôrazňujeme fakt, že tento stav je ovplyvnený aj vyššou odolnosťou metódy k-medoids voči odľahlým prvkom. Pokiaľ máme informáciu, že sa v datasete nachádzajú jedinečné objekty, ktorých existenciu nechceme zanedbať, musíme byť pri voľbe vhodnej metódy opatrní.

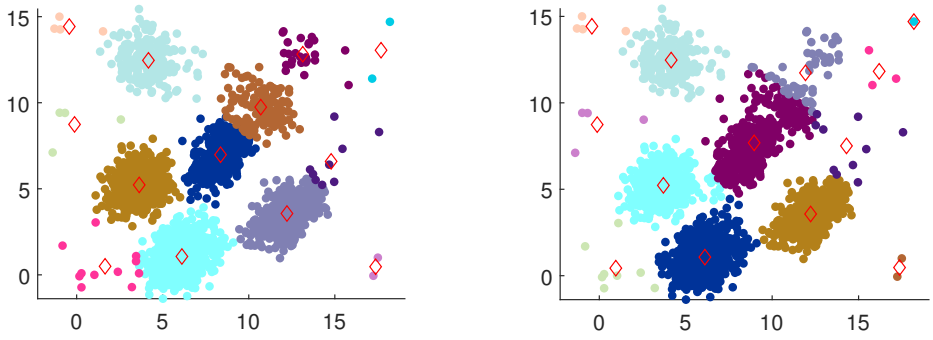
	logaritmická metrika			
Metóda	k-means	k-means++	dvojfázový k-means	k-medoids
J_{min}	0,57	0,57	0,96	0,59
J_{priem}	0,66	0,64	1,08	0,81
J_{med}	0,67	0,63	1,17	0,72
E_{max}	611,36	643,09	723,26	619,06
E_{priem}	573,25	598,41	712,36	557,62
E_{med}	573,04	599,78	710,52	555,89

Tabuľka 3.7: Tabuľka algoritmov k-means, k-means++, k-medoids a dvojfázový k-means, pre ktorý sme navýšili počet zhlukov na $k = 13$, ktoré prebehli 111 opakovaní s náhodnými inicializáciami. Obsahuje najnižšie nadobudnuté hodnoty vnútrozhlukovej miery J, maximálne nadobudnuté hodnoty medzizhlukovej miery E, priemerne dosiahnuté hodnoty vnútrozhlukovej J a medzizhlukovej E miery a ich mediány s použitím logaritmickej vzdialenosti.



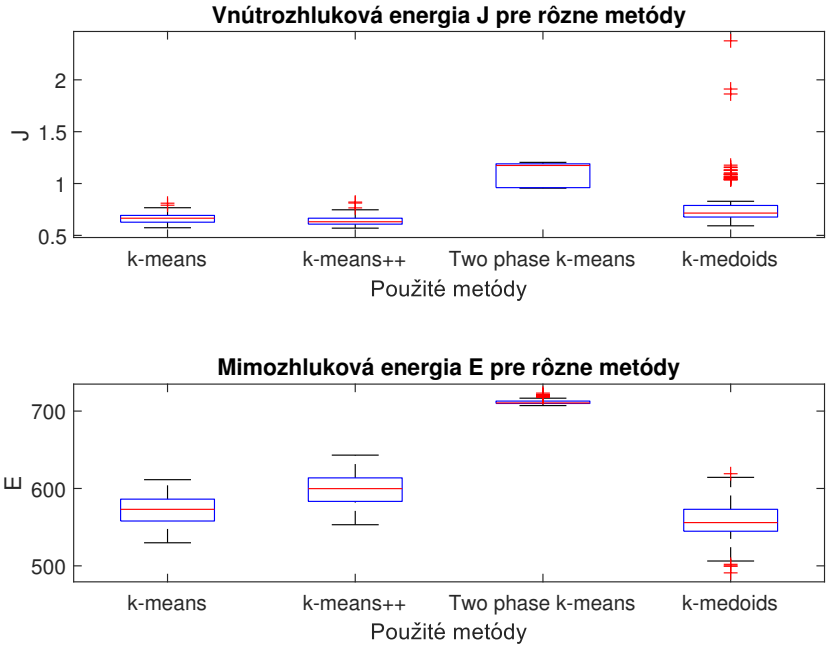
(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.32: Výsledok zhlukovacieho procesu k-medoids, $k = 13$ pre dáta so šumom podľa tabuľky 3.7.



(a) Riešenie pre J_{min} (b) Riešenie pre E_{max}

Obr. 3.33: Výsledok zhlukovacieho procesu dvojfázového k-means, $k = 13$ pre dáta so šumom podľa tabuľky 3.7.



Obr. 3.34: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 13$ v priebehu 111 behov jednotlivých algoritmov aplikovaných na dáta so šumom s náhodnými inicializáciami. Výpočty sú založené na logaritmickej vzdialenosti.

3.2.3 Obrysový graf

Na základe obrysových šírok z časti 3.1 možno určitým spôsobom vizualizovať výsledok zhlukovania obrysovým grafom. Pri tejto voľbe vizualizácie zostrojíme graf, kde na zvislú ypsilovonú os vynášame jednotlivé objekty usporiadané podľa zhlukov a hodnôt obrysových šírok, to znamená, že najprv vykreslíme prvky zhluku 1 zoradené od najvyššej obdržanej hodnoty obrysovej šírky po najnižšiu, potom rovnakým spôsobom vykreslíme zhluk 2, 3 až k.

Na vodorovnú os vynášame hodnoty obrysových šírok (vo forme vodorovných čiar spájajúcich body na osi ypsilon kde ležia objekty s bodom ich hodnoty obrysovej šírky). Pre každý zhluk nám vznikne na grafe separovaná oblasť, ktorej výška je ekvivalentná počtu objektov, ktoré sa v danom zhluku nachádzajú.

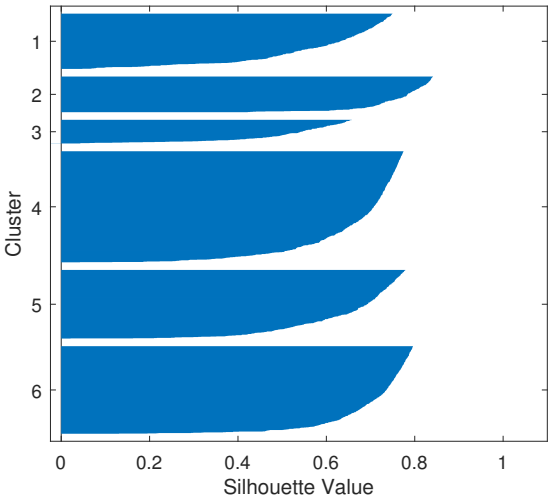
Z jednotlivých oblastí môžeme vyčítať kvalitu zaradenia objektov do zhlukov. Pokiaľ hodnoty obrysových šírok siahajú doprava k hodnote 1, ukazuje to na dobré zaradenie, pokiaľ je hodnota obrysovej šírky v okolí 0, môžeme povedať, že daný objekt je možné zaradiť aj inam (k najbližšiemu cudziemu zhluku) a ak hodnoty siahajú doľava k -1, dostávame informáciu, že daný prvok je zaradený nesprávne. Vďaka tomuto ukazateľu máme možnosť výsledok zhlukovania prehodnotiť, prípadne objekty, ktorých obrysové šírky sú blízke -1, môžme zaradiť k iným zhlukom manuálne alebo overiť, či nejde o objekty odľahlé a postaviť ich samostatne. Nespornou výhodou tejto vizualizácie riešenia je to, že nezáleží na veľkosti dimenzie príznakov a tým pádom možno zobraziť riešenie obecne n-rozmerného problému.

Ukážeme si príklady obrysových grafov. Pre porovnanie sme si zvolili riešenia s minimálnou energiou J a to konkrétne pre k-means pre $k = 6$, k-means++ pre $k = 3$, dvojfázový k-means pre $k = 7$ a $k = 12$ s použitím euklidovskej metriky a dvojfázový k-means a k-medoids pre $k = 13$ s logaritmickou metrikou.

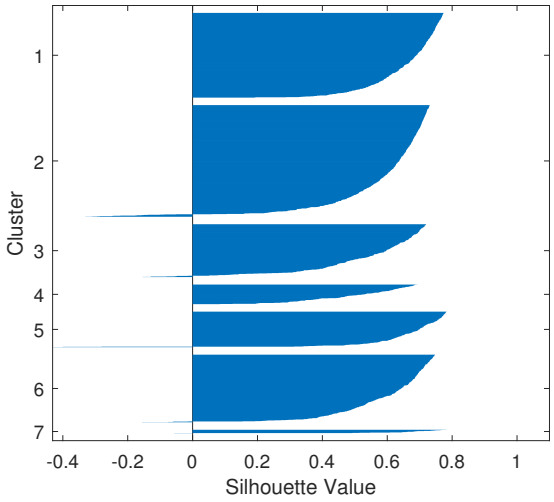
Na grafe 3.35a pre k-means, $k = 6$ s vyobrazením na 3.16a vidíme, že žiadny z objektov nemá zápornú hodnotu obrysovej šírky, naopak hodnoty obrysových šírok sú vysoké.

Graf 3.35b nám ukazuje, že aj keď nám prišlo riešenie 3.22a správne, nájdu sa objekty so zápornými hodnotami obrysových šírok, čo značí chybné zaradenie spomínaných objektov. Zároveň graf poukazuje na to, že oddelenie malého horného zhluku (na grafe znázornený ako č. 7) je správne. Každý jeho objekt dosahuje vysoké hodnoty obrysu.

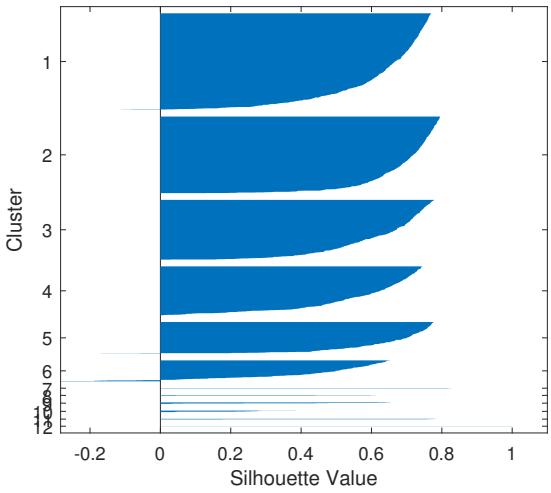
Ďalej si možno všimnúť na grafe 3.35e pre k-means++, $k = 3$, že hodnoty obrysových šírok príliš nepresahujú do záporných čísel, avšak zhluk č. 1 a zhluk č. 3 sú v porovnaní s druhým značne horšie zaradené. Rozumieme tým, že prvky týchto zhlukov dosahujú polovičné hodnoty obrysových šírok. To odpovedá samostatnému postaveniu tyrkysového zhluku na 3.8a.



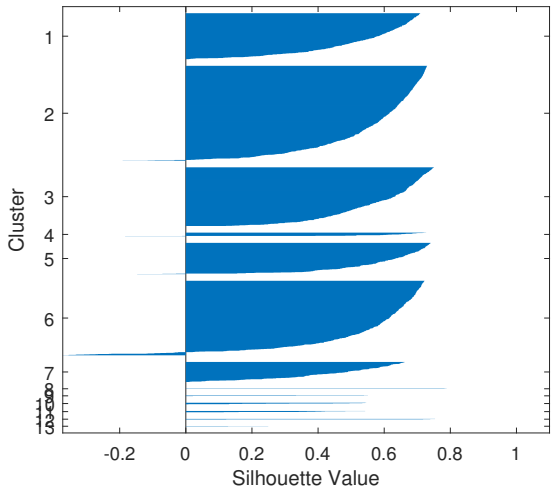
(a) Obrysový graf riešenia k-means, kde $k = 6$, s najnižšou dosiahnutou hodnotou J 3.16a



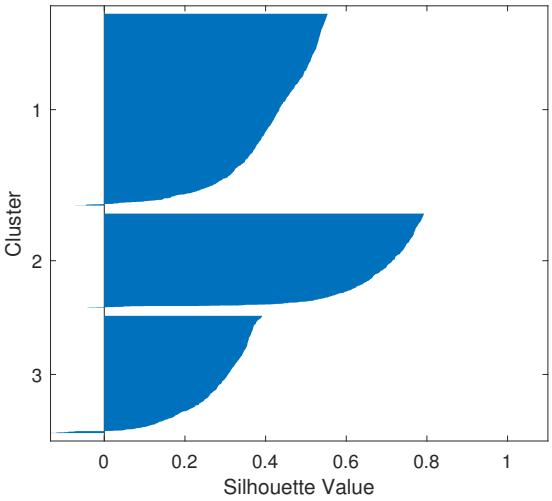
(b) Obrysový graf riešenia dvojfázového k-means, kde $k = 7$, s najnižšou dosiahnutou hodnotou J 3.22a



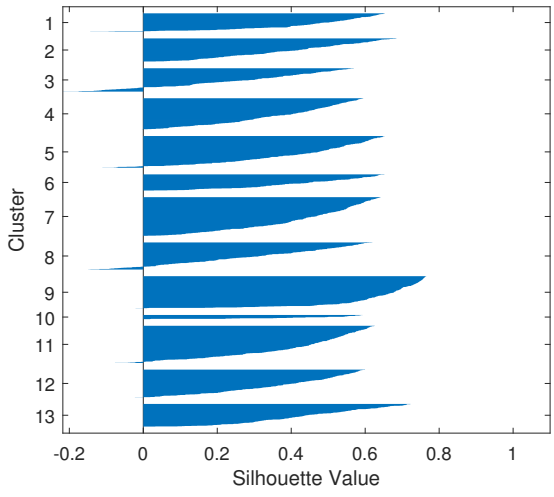
(c) Obrysový graf riešenia dvojfázového k-means, kde $k = 12$, s najnižšou dosiahnutou hodnotou J 3.30a



(d) Obrysový graf riešenia dvojfázového k-means, kde $k = 13$, s najnižšou dosiahnutou hodnotou J 3.33a



(e) Obrysový graf riešenia k-means++, kde $k = 3$, s najnižšou dosiahnutou hodnotou J 3.8a



(f) Obrysový graf riešenia k-medoids, kde $k = 13$, s najnižšou dosiahnutou hodnotou J 3.32a

3.3 Voľba energie

V tejto časti uvedieme experimenty s vnútrozhlukovými a mimozhlukovými energiami. Základ našich algoritmov popísaných v kapitole 2 tvorí vnútrozhluková suma štvorcov vzdialeností od centra. Pozrieme sa, čo by sa stalo, keby sme uvažovali iné formy energie. Upozorňujeme, že telo algoritmov sme neupravovali, len sme v každom kroku vyčísľovali inú formu J respektíve E.

Zhlukovanie stavia na idei minimalizácie vnútrozhlukovej miery J respektíve maximalizácií mimozhlukovej miery E: pokiaľ sú všetky objekty v jedinom zhluku, je táto miera J najväčšia respektíve E je nulová, na druhej strane, keď každý objekt stojí samostatne je J nulová respektíve E je maximálna. Na základe tohto princípu kladieme dôraz, aby J klesala a E rástla (vrámci iterácií ustálenia pre konkrétne k, aj v rámci rastúceho počtu zhlukov). Nové energie sme vyobrazili v grafoch s použitím funkcie k-means, kde $k = 6$, s rovnakou inicializáciou pre každý prípad a taktiež sme vykreslili vývoj oboch druhov energií v intervale $k = 1$ až $k = 17$.

Pre experimenty sme zostavili sedem vnútrozhlukových a sedem mimozhlukových mier. Ako prvú variantu vnútrozhlukovej účelovej funkcie J sme zvolili normovanú sumu štvorcov vzdialeností na jeden prvok podľa (2.2):

$$J = \frac{WCSS}{n} = \frac{1}{n} \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} d_E^2(\mathbf{x}_i, \mathbf{c}_h), \tag{3.6}$$

a mimozhlukovú mieru volíme na základe metódy priemernej väzby (1.11)

$$E = \sum_{r=1} \sum_{q=1} d_{rq}. \tag{3.7}$$

Druhou kombináciou je vnútrozhluková miera vo forme priemernej vzdialenosti prvkov k svojim centrám

$$J = \sum_{h=1}^{k} \frac{1}{n_h} \sum_{\mathbf{x} \in \mathcal{C}_h} d(\mathbf{x}, \mathbf{c}_h), \tag{3.8}$$

kde n_h je počet prvkov zhluku $\mathcal{C}_h$. Mimozhlukovú mieru E počítame podľa metódy najbližšieho suseda (1.9)

$$E = \sum_{r=1} \sum_{q=1} d_{rq}. \tag{3.9}$$

Ďalšia experimentálna funkcia J predstavuje maximálnu šírku obalu zhluku okolo svojho centra

$$J = \sum_{h=1}^{k} [\max_{\mathbf{x} \in \mathcal{C}_h} d(\mathbf{x}, \mathbf{c}_h) - \min_{\mathbf{x} \in \mathcal{C}_h} d(\mathbf{c}, \mathbf{c}_h)], \tag{3.10}$$

predstavuje teda sumu rozdielov vzdialeností najvzdialenejšieho prvku a centra s najbližším prvkom a centrom. E počítame podľa metódy najvzdialenejšieho suseda (1.10). Štvrtý experiment zahŕňa J ako súčet vzdialeností medzi centrami a ich najbližšie ležiacimi prvkami, násobený počtom prvkov v zhluku

$$J = \sum_{h=1}^{k} n_h \min_{\mathbf{x} \in \mathcal{C}_h} d(\mathbf{x}, \mathbf{c}_h), \tag{3.11}$$

a E ako súčet jednotlivých vzdialeností dvoch centier

$$E = \sum_{r=1}^{k} \sum_{q=1}^{k} d(\mathbf{c}_r, \mathbf{c}_q). \tag{3.12}$$

V ďalšoum pokuse J predstavuje súčet vzdialeností medzi centrami a ich najďalej ležiacimi prvkami

$$J = \sum_{h=1}^{k} \max_{\mathbf{x} \in \mathcal{C}_h} d(\mathbf{x}, \mathbf{c}_h) \tag{3.13}$$

a E sme sformulovali ako rozdiel vnútrozhlukových energií WCSS systému kedy $k = 1$ a systému v súčasnom stave

$$E = WCSS(1) - WCSS(k) = \sum_{\mathbf{x}_i \in C_1} d_E^2(\mathbf{x}_i, \mathbf{c}_1) - \sum_{h=1}^{k} \sum_{\mathbf{x}_i \in C_h} d_E^2(\mathbf{x}_i, \mathbf{c}_h). \tag{3.14}$$

Predposlednou voľbou je tvar vnútrozhlukovej miery v podobe súčtu mediánov vzdialeností centra a prvku

$$J = \sum_{h=1}^{k} \operatorname*{med}_{\mathbf{x} \in \mathcal{C}_h} d(\mathbf{x}, \mathbf{c}_h) \tag{3.15}$$

a E počítame ako rozdiel plochy resp. objemu, ktorú zaberajú objekty v priestore (v našom prípade ako obdĺžnik) a plochy resp. objemu, ktoré zaberajú zhluky, (počítame ako obsah kruhu so stredom v centroide a polomerom s hodnotou vzdialenosti medzi centrom a priemerne vzdialeným prvkom). Symbolicky teda môžeme písať:

$$E = V(\mathbf{X}) - \sum_{h=1}^{k} \overline{V}(\mathcal{C}_h), \tag{3.16}$$

kde pod $V(\mathbf{X})$ rozumieme objem, ktorý zaberajú všetky objekty a $\overline{V}(\mathcal{C}_h)$ značí objem zhluku $\mathcal{C}_h$ v závislosti na priemernom prvku. Posledný, siedmy, experiment je založený na vnútrozhlukovej miere pozostávajúcej zo súčtu priemerov jednotlivých zhlukov

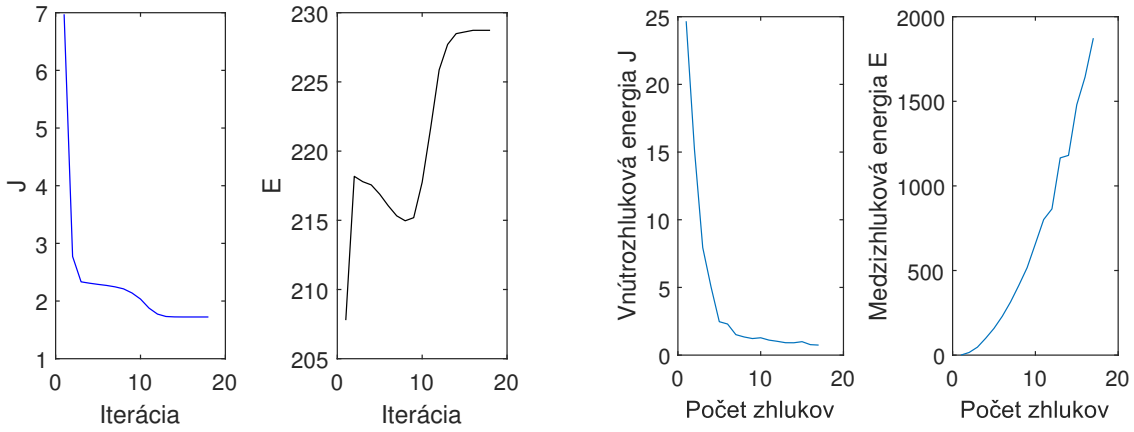
$$J = \sum_{h=1}^{k} diam(\mathcal{C}_h) = \sum_{h=1}^{k} [\max_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_h} d(\mathbf{x}_i, \mathbf{x}_j)] \tag{3.17}$$

a E predstavuje rozdiel energií pre metódu najvzdialenejšieho suseda a metódu najbližšieho suseda:

$$E = E_{far} - E_{near} = \sum_{r=1} \sum_{q=1} \{ \min_{\substack{\mathbf{x}_i \in \mathcal{C}_r \\ \mathbf{x}_j \in \mathcal{C}_q}} (d(\mathbf{x}_i, \mathbf{x}_j)) \} - \sum_{r=1} \sum_{q=1} \{ \max_{\substack{\mathbf{x}_i \in \mathcal{C}_r \\ \mathbf{x}_j \in \mathcal{C}_q}} (d(\mathbf{x}_i, \mathbf{x}_j)) \}. \tag{3.18}$$

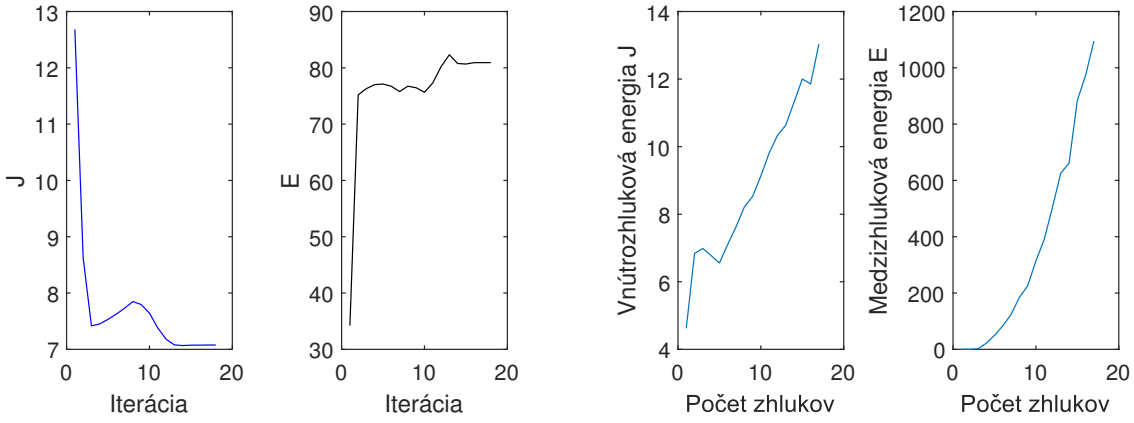
Z grafov môžeme vidieť, že naše požiadavky na energie J spĺňa: normovaná WCSS 3.36, pričom energia J priemerne vzdialených prvkov 3.37, energia šírky obalu 3.38 a energia počítaná pomocou priemeru okolo centra 3.42 síce spĺňajú požiadavku na pokles v priebehu cyklu pre pevné k, ale nespĺňajú pokles s rastúcim k. U energie spojenej s mediánmi môžme váhať vrámci vývoja energie a použiť graf 3.41a na určenie počtu zhlukov $k = 6$ či $k = 7$ v metóde kolena. Všimnime si, že určité koleno sa pre $k = 7$ nachádza aj na grafe energie E. Naše požiadavky na mimozhlukovú energiu E spĺňajú všetky energie až na E najvzdialenejšieho suseda, kedy v metóde k-means pre $k = 6$ pozorujeme rastúci trend až pre vyššie iterácie. Normovaná suma štvorcov vzdialeností teda zostáva najvhodnejšou voľbou.

Z tohto experimentu možno usúdiť, že bez zásahu do princípu metódy nemožno používať ľubovoľný tvar energie, čo vyplýva najmä z (2.3), kde vidíme, že stred zhluku ako priemer prvkov je určený z minimalizácie WCSS (2.2).



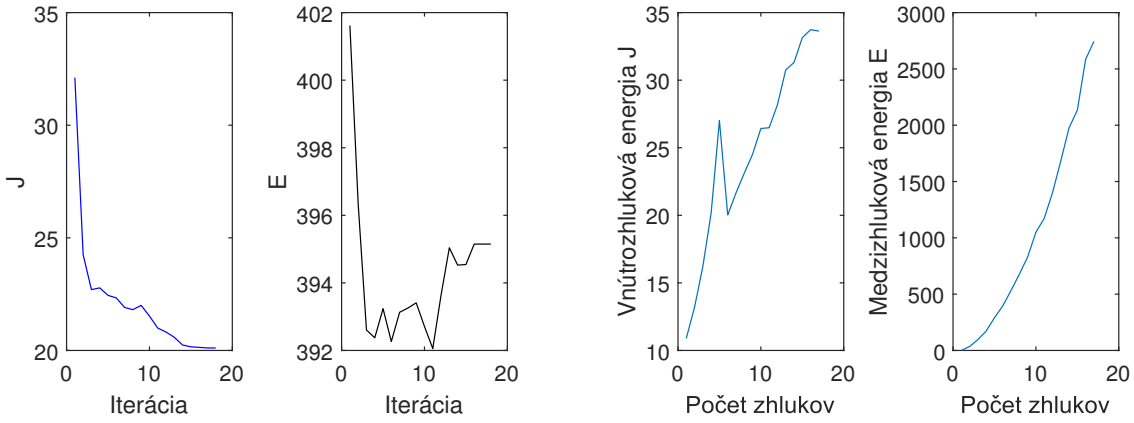
(a) (b)

Obr. 3.36: Znázornenie priebehu energií J (3.6) a E (3.7) počas jedného behu k-means algoritmu na umelé dáta bez šumu pre $k = 6$ (a). Vývoj energií J a E pre funkciu k-means pre $k = 1 \ldots 17$ (b).



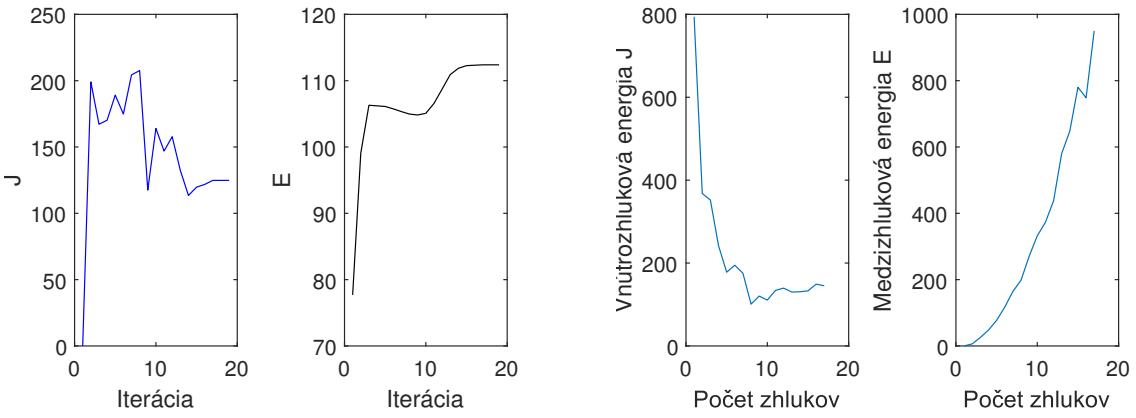
(a) (b)

Obr. 3.37: Znázornenie priebehu energií J (3.8) a E (3.9) počas jedného behu k-means algoritmu na umelé dáta bez šumu pre $k = 6$ (a). Vývoj energií J a E pre funkciu k-means pre $k = 1 \ldots 17$ (b).



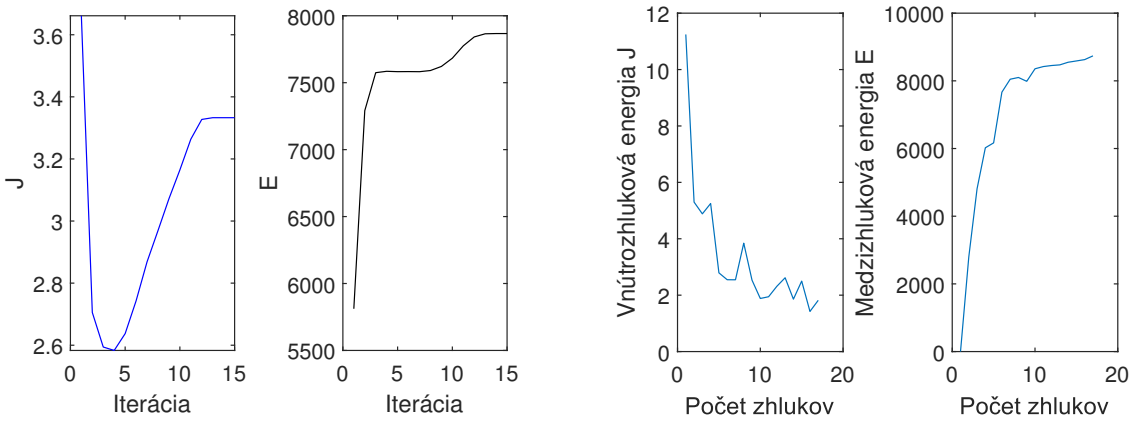
(a) (b)

Obr. 3.38: Znázornenie priebehu energií J (3.10) a E (1.10) počas jedného behu k-means algoritmu n aumelé dáta bez šumu pre $k = 6$ (a). Vývoj energií J a E pre funkciu k-means pre $k = 1 \ldots 17$ (b).



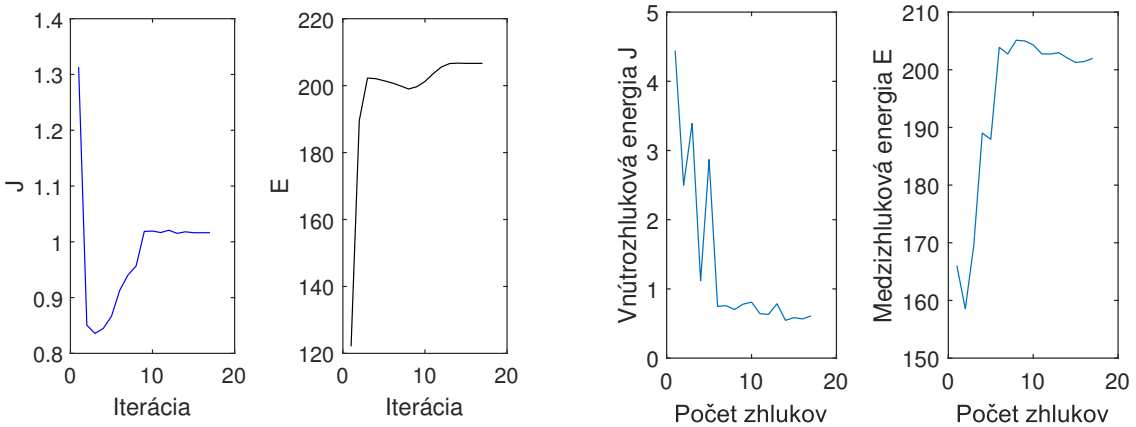
(a) (b)

Obr. 3.39: Znázornenie priebehu energií J (3.11) a E (3.12) počas jedného behu k-means algoritmu na umelé dáta bez šumu pre $k = 6$ (a). Vývoj energií J a E pre funkciu k-means pre $k = 1 \ldots 17$ (b).



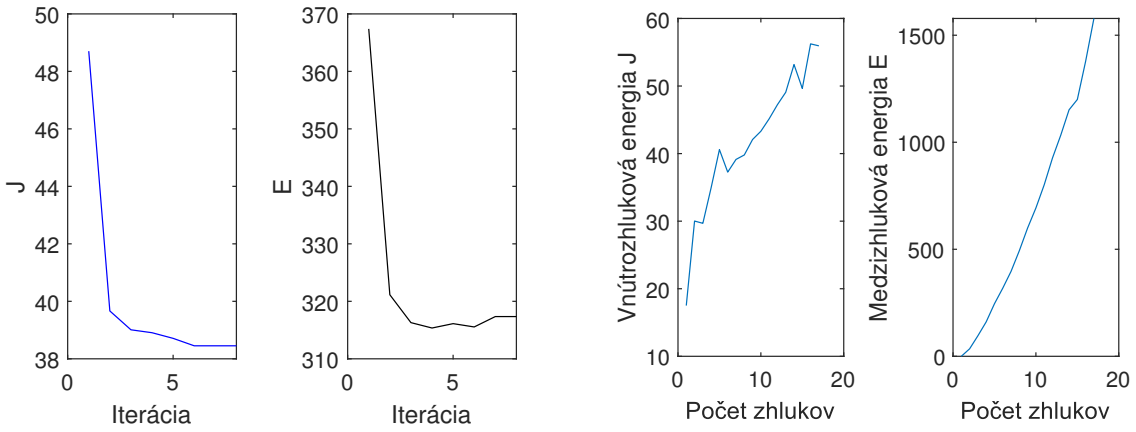
(a) (b)

Obr. 3.40: Znázornenie priebehu energií J (3.13) a E (3.14) počas jedného behu k-means algoritmu na umelé dáta bez šumu pre $k = 6$ (a). Vývoj energií J a E pre funkciu k-means pre $k = 1 \ldots 17$ (b).



(a) (b)

Obr. 3.41: Znázornenie priebehu energií J (3.15) a E (3.16) počas jedného behu k-means algoritmu na umelé dáta bez šumu pre $k = 6$ (a). Vývoj energií J a E pre funkciu k-means pre $k = 1 \ldots 17$ (b).



(a) (b)

Obr. 3.42: Znázornenie priebehu energií J (3.17) a E (3.18) počas jedného behu k-means algoritmu na umelé dáta bez šumu pre $k = 6$ (a). Vývoj energií J a E pre funkciu k-means pre $k = 1 \ldots 17$ (b).

3.4 Zhrnutie

V tejto kapitole sme učinili niekoľko experimentov na umelých dvojrozmerných dátach. Môžeme si na modeli bez šumu všimnúť, že ak máme väčší dataset s rôznou koncentráciou objektov, pri opakovanom použití metód (my sme použili stojedenásť opakovaní), metódy k-means, k-means++ a k-medoids takmer splývajú v zmysle hodnoty dosiahnutých energií.

Ak máme informáciu, že v dátach nie sú odľahlé prvky, je zbytočné použiť dvojfázový k-means z dôvodu značne dlhšieho výpočetného času a vyššej hodnoty minimálnej dosiahnutej vnútrozhlukovej energie.

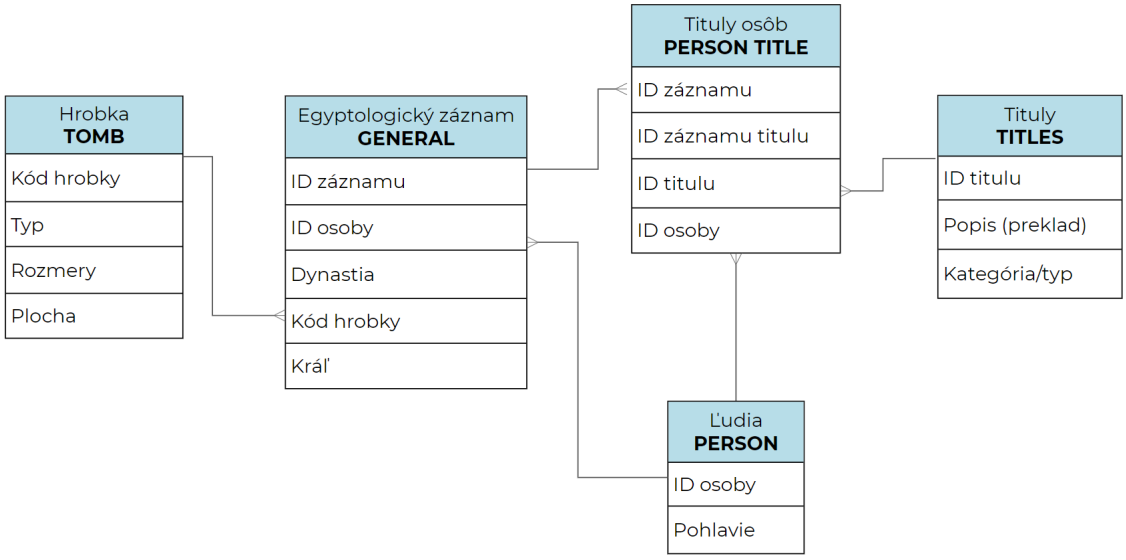
Ak však máme podozrenie, že v dátach sú odľahlé prvky alebo šum, je výhodnejšie použiť dvojfázový k-means. Na modeli dát so šumom môžeme vidieť, že táto dvojfázová metóda nadobúda energeticky značne lepšie výsledky pre separátne určený vhodný počet zhlukov. V prípade, že stanovíme aj pre klasické metódy rovnaký počet zhlukov (v našom prípade 13) je vidieť, že tieto metódy ani zďaleka nereagujú adekvátne: v rámci znižovania vnútrozhlukovej energie dochádza k deleniu mnohopočetných zhlukov na úkor izolácie vzdialených objektov. Dvojfázový k-means na úkor vyššej hodnoty energie dokáže tieto odľahlé prvky oddeliť.

Na záver sme uskutočnili experimenty so zmenami predpisov energií. Zistili sme, že ak ponecháme pôvodný princíp algoritmov, tzn. sú zostrojené tak, aby minimalizovali vhodnú formu WCSS (2.2) a (2.17), potom nie je vhodné energie ľubovoľne zamieňať. V kapitole 2 sme ukázali (2.3), že metóda k-priemerov má základ v tom, že práve priemer minimalizuje WCSS energiu.

Kapitola 4

Aplikácia zhlukovej analýzy na egyptologické dáta

Ústredný motív tejto kapitoly spočíva v analýze dát získaných Českým egyptologickým ústavom Univerzity Karlovej. Českí egyptológovia v priebehu desiatok rokov archeologických vykopávok a výskumov zaznamenávali údaje z priečelí objavených hrobiek. Údaje pozostávajú z informácií o osobách ako napríklad konkrétne meno, obdobie života a meno panovníka, ktorý vládol počas ich života, pohlavie, tituly, ktoré počas života nadobudol majiteľ hrobky, dokonca aj jeho príbuzenstvo, priatelia či služobníctvo. Postupne sa na základe týchto záznamov vytvára a neustále dopĺňa (napríklad o informácie o type hrobky a jej charakteristikách) obsiahla databáza nazývaná *Maat-base*.



Obr. 4.1: Schéma obdržaných dát z *Maat base*

Pre potreby tejto práce sme obdržali časť týchto údajov v dvoch etapách a to z dôvodu, že egyptologický projekt sa neustále vyvíja a údaje sa obmieňajú a dopĺňajú. V prvej etape sme obdržali dataset 4197 záznamov ľudí žijúcich počas piatej a šiestej periódy respektíve dynastie (piata skorá *5. early*, prostredná *5. middle*, neskorá *5. late* a šiesta skorá *6. early*, prostredná

6. middle a neskorá *6. late* perióda) spolu s ich nadobudnutými titulmi a prevodnou tabuľkou do kategórí titulov. Tento dataset budeme spracovávať v časti 4.1.

V druhej etape sme sa oboznámili s komplexnejším súborom dát. Túto subdatabázu pre prehľad znázorníme schémou 4.1. Tieto dáta tvorí niekoľko tabuliek, ktoré na seba nadväzujú. Základ nášho systému tvorí tabuľka egyptologických záznamov **General**, zložená z čísla záznamu v *Maat-base*, čísla osoby, ktoré zároveň považujeme za charakteristiku totožnú s menom osoby, dynastiu života, kódu hrobky, z ktorej sa uvedené údaje zaznamenali a mena kráľa. Túto tabuľku prepájame s ostatnými tabuľkami. Prvou je tabuľka o charakteristikách hrobky **Tomb**, pričom sa jedna hrobka môže spájať s viacerými záznamami. Ďalej máme k dispozícií tabuľku **Person** - zoznam ľudí doplnený o informáciu o ich pohlaví a tabuľku údajov o tituloch osôb **Person Title**, ktorá prepája osobu a všetky jej tituly naprieč záznamami. Posledným článkom nášho systému je tabuľka **Titles** - popisu titulov a ich zaradenia do globálnych kategórií.

Pohlavie	Počet ľudí
žena	812
muž	3518
muž(?)	1
neisté	3
neuvedené	4

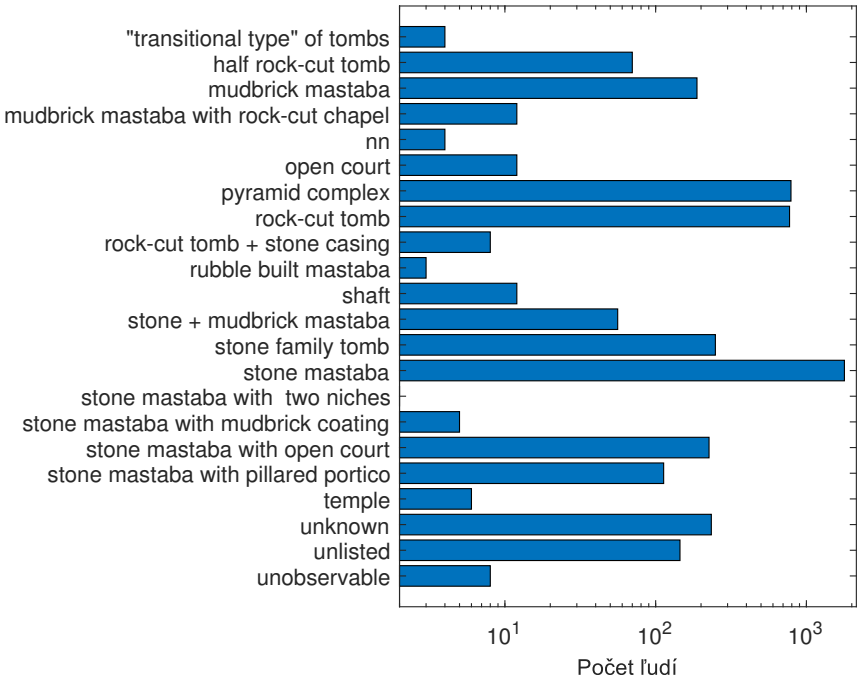
Tabuľka 4.1: Tabuľka pohlaví a počtu osôb príslušného pohlavia podľa obdržaných dát z *Maat base*

Za účelom aplikácie metód zhlukovej analýzy si potrebujeme zvoliť príznakový priestor a charakterizovať objekty. Našimi objektmi budú konkrétni ľudia pre prvý dataset resp. konkrétne hrobky pre dataset druhý, teda objekt človeka resp. hrobky možno popísať príznakmi - údajmi, ktoré máme k dispozícií - teda napríklad za príznaky volíme pohlavie, typ hrobky, titul 1 (resp. kategóriu 1), titul 2 atď..

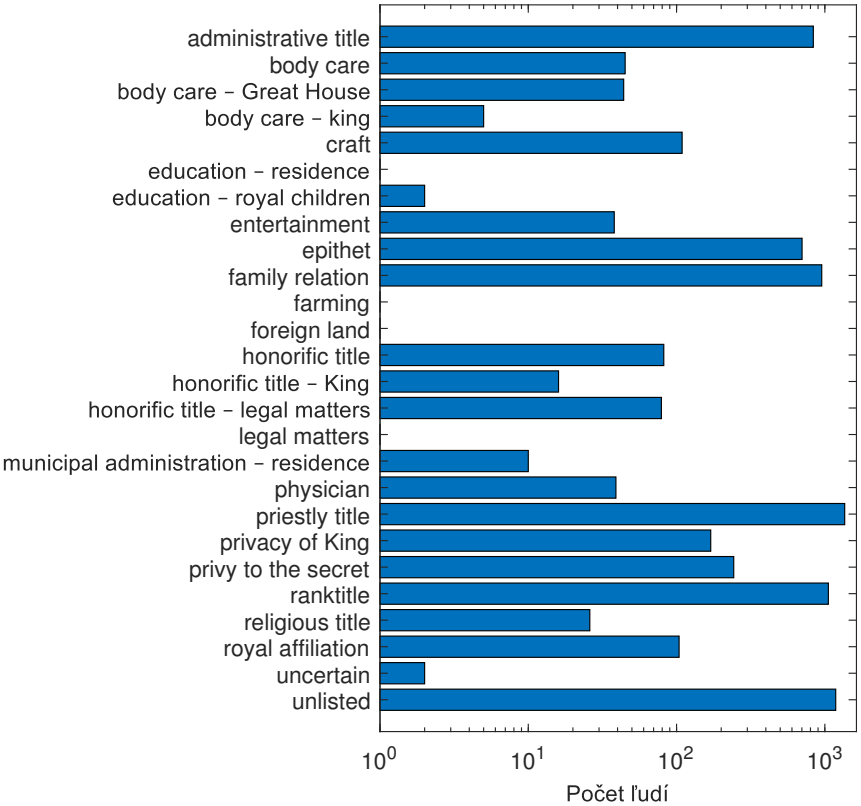
Pre bližšie pochopenie obdržaných dát z druhej etapy ukážeme zastúpenie objektov vrámci príznakov. Upozorňujeme, že názvy kategórií titulov, ani typu hrobiek neprekladáme z dôvodu konzistentnosti s inými prácami s egyptologickou databázou *Maat-base*. Uvádzame početnosti jednotlivých pohlaví vrámci našich dát v tabuľke 4.1. Náš dataset tvorí 812 žien, 3518 mužov, 1 asi muž a o 3 a 4 ľuďoch si nie sme istý či nebolo uvedené ich pohlavie.

Ďalej vizualizujeme na histogramoch zastúpenia zmienok osôb v daných typoch hrobiek na obrázku 4.2, to jest napríklad 1787 zmienok bolo v hrobkách typu *stone mastaba*, 790 a 773 zmienok v hrobkách typu *pyramid complex* a *rock-cut tomb*, najmenej zmienok náleží hrobkám *stone mastaba with two niches*, *rubble built mastaba*, *transitional type of tombs* a *nn*, kde sa nachádzajú po rade 3, 4, 4 a 4 zmienky.

Na grafe 4.3 ukazujeme koľko ľudí bolo držiteľom titulov z jednotlivých uvedených kategórií, pričom pre zjednodušenie sme zlúčili niekoľko poddruhov administratívnych kategórií titulov do jednej *administrative title* a 836 ľudí disponovalo titulom, ktorý možno zaradiť do tejto kategórie, rovnako sme takým spôsobom zložili kategóriu *priestly title*, kam spadá 1361 ľudí, zlúčili sme do jednej špecifické kategórie *epithet* so 702 ľuďmi, zoskupili sme podkategórie *privacy of King* (170 ľudí), *rank title* (1056 ľudí), niekoľko zameraní do obecnej kategórie *craft* (109), *entertainment* (38) či *privy to the secret* (243) a *physician* (39). Najmenej ľudí disponovalo titulom z kategórií *education - residence*, *farming*, *forein land*, *legal matters*, to jest po 1 človeku. Poznamenajme, že pre nerovnomerné početnosti volíme pre znázornenie na oboch histogramoch logaritmickú škálu v osi x.



Obr. 4.2: Histogram počtu ľudí(záznamov) u konkrétnych typov hrobiek v súbore General obdržaných dát z *Maat base*. Os x je na logaritmickej škále.



Obr. 4.3: Histogram počtu ľudí na nadobudnutý titul z konkrétnych kategórií z obdržaných dát z *Maat base*. Os x je na logaritmickej škále.

Ďalšou náplňou v tejto kapitole je testovanie základnej podoby zhlukovacích algoritmov na obdržanom datasete a čiastočnej interpretácií výsledkov použitých metód, ktoré však nemusia byť pre odborníkov dostatočne valídne, avšak pre účely tejto bakalárskej práce sú v rámci testovania metód postačujúce. Našou primárnou úlohou je analyzovať tituly staroegyptskej spoločnosti.

Analytickú časť kapitoly rozdelíme na dva oddiely podľa obdržaných dát, teda na oddiel náležiaci základnému balíku záznamov: *človek↔titul↔obdobie života* a na oddiel rozšíreného datasetu o *pohlavie*, *kód krobky* a *typ hrobky*. V každej časti vytvoríme dátovú maticu $\mathbf{X}$ zloženú z dostupných egyptologických záznamov, aplikujeme na ňu vybrané metódy zhlukovej analýzy: k-means a k-modes. Predstavíme výsledky vzniknuté použitím týchto metód.

4.1 Základný dataset

Našou úlohou v tejto sekcií je pokúsiť sa detekovať skupiny na základe titulov, ktoré daným osobám náležali, s využitím teórie zhlukovania na prvom základnom datasete. Je známych 1966 titulov, ktoré spadajú do 69 základných kategórií, a zároveň konkrétny objekt mohol disponovať aj viacerými titulmi zároveň. Poznamenajme, že jeden titul môže spadať do jednej alebo viac kategórií titulov. Zoznam kategórií titulov pre túto úlohu etapy 1 nájdeme v prílohe 4.2.3 ako tabuľlu 4.13.

4.1.1 Predspracovanie dát

Prvým krokom je tvorba vstupnej matice $\mathbf{X}$ algoritmov. Rozhodli sme sa využiť zúžený príznakový priestor na 69 kategórií titulov, z toho dôvodu vytvoríme vstupnú maticu, ktorej stĺpce (príznaky) odpovedajú kategórii titulu, ktorú bolo možné nadobudnúť. Stavom kedy objekt nadobúda určitú hodnotu danej črty rozumieme, že Egypťan mal respektíve nemal danú kategóriu titulu. Tejto logike prislúcha binárny systém hodnôt teda 1 respektíve 0.

$\mathbf{X}$	Kategória 1	Kategória 2	$\cdots$	Kategória 69
	$\{0;1\}$	$\{0;1\}$	$\cdots$	$\{0;1\}$

Druhým krokom, ktorý zjednoduší ďalšiu prácu, je vyčlenenie takých objektov, ktorých hodnoty všetkých príznakov sú nulové. Z týchto nulových vektorov objektov vytvoríme samostatný zhluk: *Egypťania bez akejkoľvek kategórie*. Na ďalšie spracovanie teda máme objekty s aspoň jednou nenulovou črtou. Za účelom porovnávania vytvoríme tri podmnožiny: objekty z piatej periódy, objekty zo šiestej periódy a celkový dataset oboch periód dohromady. Toto rozdelenie nám môže umožniť porovnať vývoj medzi piatou a šiestou érou ríše.

V našom datasete sa nachádza 2968 ľudí, ktorí žili počas 5. periódy, z toho 2343 Egypťanov malo aspoň jeden titul a 625 ľudí tvorí skupinu bez akéhokoľvek titulu, ktorý by spadal do nejakej zo 69 kategórií. Z obdobia 6. periódy v našom datasete nachádza 1540 ľudí, z toho 1201 malo aspoň jednu zastúpenú kategóriu titulu a 339 jedincov nedisponovalo žiadnou kategóriou titulu. Ak skúmame obe periódy naraz, potom v nich skúmame 4197 Egypťanov, z ktorých 3353 disponovali aspoň jedným z kategorizovaných titulov a 844 ľudí celkovo nemalo nijaký titul.

4.1.2 Aplikácia zhlukovacích metód

Za zhlukovaciu metódu sme zvolili metódu k-modes, popísanú algoritmom 9, ktorá pracuje s primitívnou metrikou (ak majú objekty rovnakú kategóriu titulu, ich vzdialenosť je v danej súradnici 0, naopak, ak majú rôzne kategórie, vzdialenosť v danej dimenzii je 1, celková vzdialenosť dvoch objektov je vysčítaná cez jednotlivé dimenzie).

Ďalej sme pre porovnanie zvolili aj metódu k-means, pre ktorú použijeme blokovú metriku. Metóda k-modes pracujúca s primitívnou metrikou má súradnice centier celočíselné, teda hovorí, že daný príznak v zhluku buď majoritne bol alebo nebol zastúpený. Na dôvažok metóda k-means s blokovou metrikou na binárnom príznakovom priestore môže mať súradnice centra neceločíselné a z intervalu $(0, 1)$, pričom každá súradnica vyjadruje podiel zastúpení daného príznaku v zhluku.

Budeme postupovať po periódach (5.,6., obe), pričom zakaždým zvolený algoritmus spustíme pre rôzne hodnoty počtu zhlukov k v sto jedenástich opakovaniach a zapamätáme si také riešenie, ktoré dosiahlo najlepšiu vnútrozhlukovú energiu J. Pri nižšom počte opakovaní častejšie dochádzalo k skokovým nárastom vo funkcii J a poklesom vo funkcii E. Následne vykreslíme priebehy vnútrozhlukových energií J v rámci metódy kolena popísanev v časti 3.1, graf obrysových koeficientov zo sekcie 3.1 a na ich základe sa pokúsime určiť hodnotu počtu zhlukov k.

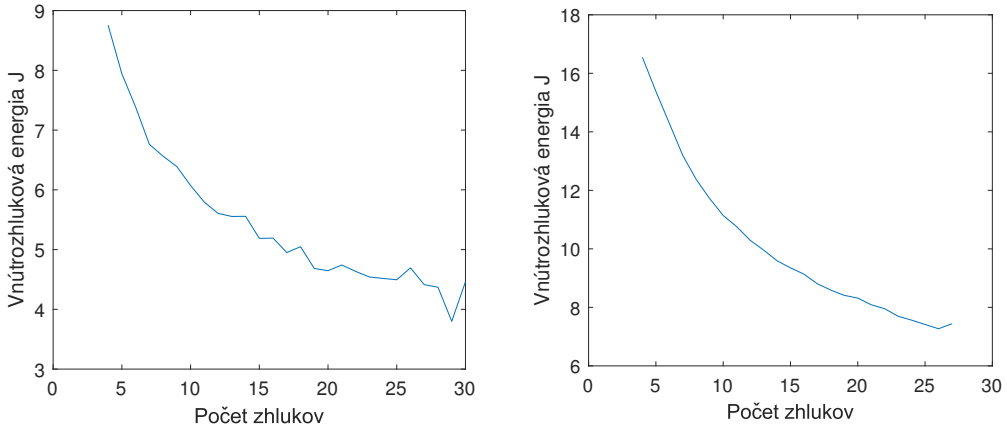
Poznamenajme, že algoritmy sme oproti pôvodnému popisu v kapitole 2 upravili poistkou voči vzniku prázdnych zhlukov tak, že riešenie v ktorom sa nachádzal prázdny zhluk sme neprijali ako valídne (nedbajúc na jeho hodnotu energií). Učinili sme tak na základe pozorovania vzniku prázdnych zhlukov, ktoré ovplyvňovali hodnotu obrysového koeficientu a zároveň tým porušovali princíp zhlukovania, že *žiaden zhluk nie je prázdny*, ako sme uviedli v kapitole 1.

Na obrázkoch 4.4 až 4.12 uvádzame výsledky aplikácie zvolených algoritmov na vyššie popísané datasety. Na obrázku 4.4 je zobrazený vývoj vnútrozhlukovej energie J pre k-modes a k-means metódy, kedy sme volili $k = 4 \dots 30$ na 5. periódu. Na týchto priebehoch možno pozorovať kolísanie hodnôt pri použití metódy k-modes s primitívnou metrikou obr.:4.4a. Algoritmus sa najmä pre vyššie hodnoty k (15 a viac) zacykloval a dochádzalo k vyprázdneniu zhlukov a teda sme v stojedenástich opakovaniach nedosiahli očakávané zníženie hodnoty J pri prechode $k = h \to k = h + 1$. Energia J pre metódu k-means na seba nadväzovala plynulejšie obr.:4.4b, avšak ani na jeden z týchto grafov nemožno úspešne aplikovať metódu kolena, pretože na krivkách žiadne *plató* nepozorujeme. Identické správanie vnútrozhlukovej energie je badateľné aj pri použití metód na 6. periódu, viď graf4.7 a taktiež na obe periódy súčasne, viď graf 4.10.

Po neúspešnom pokuse využiť metódu kolena pokračujeme metódou obrysového koeficientu, ktorú na tri datasety aplikujeme rovnako: pre $k = 4 \dots 30$ s tým, že akceptujeme hodnoty obrysových šírok spolu s hodnotou J a neprijímame také riešenia, pre ktoré evidujeme prázdny zhluk. Na grafoch obrysových koeficientov pre jednotlivé periódy síce pozorujeme výskyt maxima, avšak v porovnaní so simulovaným datasetom na obr.:3.4 toto maximum nedosahuje ani polovicu maximálne možnej hodnoty, dokonca ho nemožno považovať za jednoznačne výrazne odlišiteľné.
Na grafe 4.5b chýba obrysový koeficient pre $k = 28$ a $k = 29$, pretože pre tieto hodnoty počtu zhlukov algoritmus v ani jednom z opakovaní nenašiel plný počet zhlukov. Aj napriek tomu, že k určené na základe metódy obrysového koeficientu neevokuje plne dôveryhodnosť, toto k prijímame ako vhodný počet zhlukov, pretože metóda kolena nám žiadnu prijateľnú hodnotu nedala.

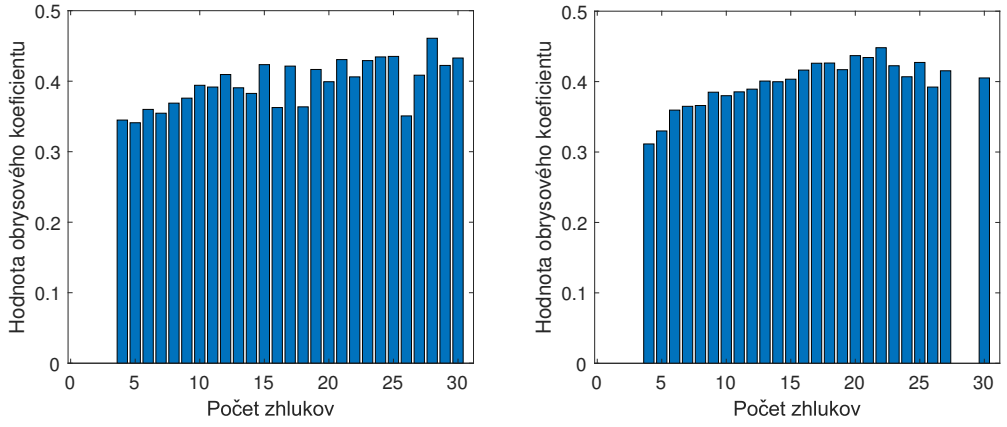
V každej perióde sme pre prijatú hodnotu k vyobrazili pomocou obrysového grafu také riešenia metód k-modes a k-means, ktoré dosiahli v rámci opakovaní najnižšiu hodnotu vnútrozhlukovej energie J. Na týchto grafoch 4.6, 4.9, 4.12 možno pozorovať, že algoritmus k-means ap-

likovaný na dáta z každej periódy nenadobúdal toľko záporných hodnôt ako algoritmus k-modes pre rovnaké dáta. Na druhej strane aj napriek výskytu záporných hodnôt obrysových šírok pozorujeme, že hlavné zhluky (viacpočetné) dosahovali vyššie hodnoty, dokonca rozoznávame na 4.6a zhluk 23, ktorý má všetky prvky zaradené tak, že dosahujú hodnotu obrysových šírok $s(k) = 0,98$. Ten istý zhluk sa nachádza aj v šiestej perióde ako zhluk 7 s obrysovou šírkou $s(k) = 0,95$ a na grafe oboch períód nadobúda $s(k) = 0,91$ ako zhluk 28. Poznamenajme, že hodnoty k pre každú z metód boli rôzne.



(a) k-modes (b) k-means

Obr. 4.4: Priebeh vnútrozhlukovej energie J pre metódu k-modes (a), metódu k-means (b), pre $k = 4 \ldots 30$ pri aplikácií na 5. periódu.

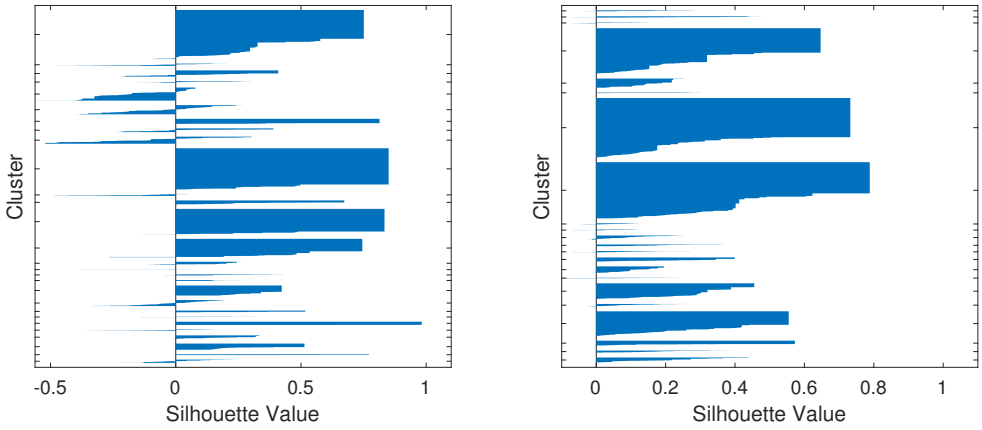


(a) k-modes (b) k-means

Obr. 4.5: Priebeh obrysových koeficientov pre metódu k-modes (a), metódu k-means (b), pre $k = 4 \ldots 30$ pri aplikácií na 5. periódu.

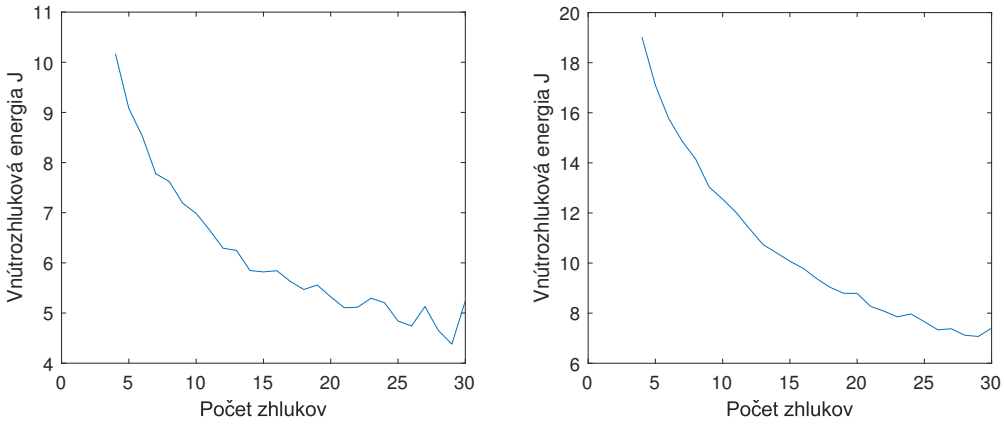
4.1.3 Vyhodnotenie

V tejto časti porovnáme výsledky algoritmov k-modes, $k = 28$ a k-means, $k = 29$, ktoré vznikli aplikáciou na objekty pre dataset celého obdobia. V tabuľke 4.2, ktorá plne odpovedá riešeniu vyobrazenému na obrysovom grafe 4.12a, je zaznamenaný prehľad všetkých zhlukov, počet ich objektov, priemernú a maximálnu vzdialenosť jedincov od svojho centra v podobe módu. V poslednom stĺpci uvádzame čísla kategórií titulov, ktoré získalo najmenej 85% objektov



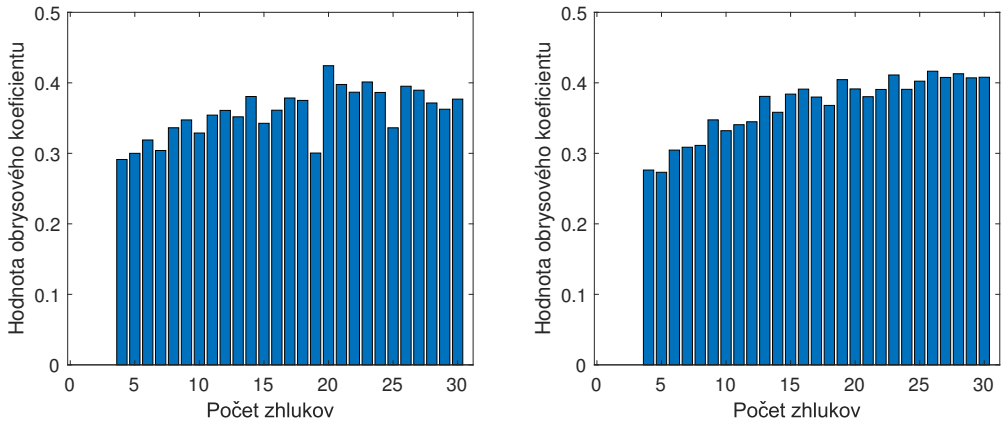
(a) k-modes (b) k-means

Obr. 4.6: Riešenie zhlukovania 5. periódy pre k-modes, $k = 28$ (a) s hodnotou obrysového koeficientu $\overline{s}(k) = 0,46$ a pre k-means, $k = 22$ (b) s hodnotou obrysového koeficientu $\overline{s}(k) = 0,45$.



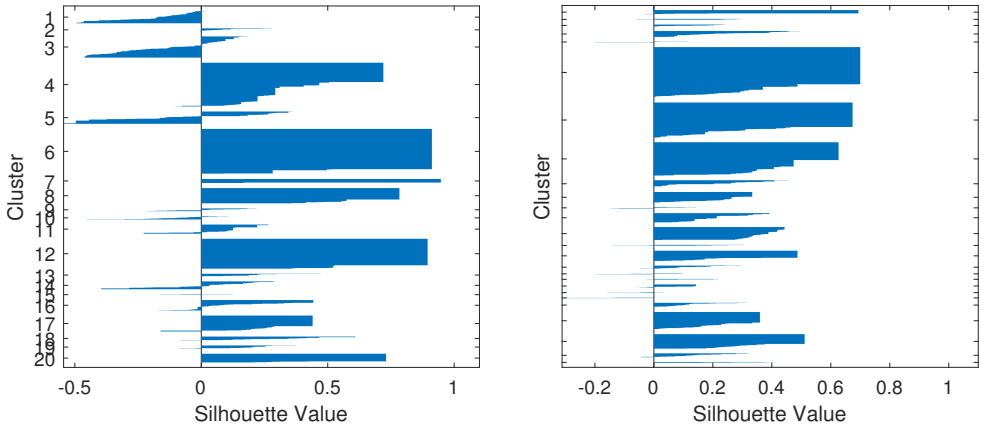
(a) k-modes (b) k-means

Obr. 4.7: Priebeh vnútrozhlukovej energie J pre metódu k-modes (a), metódu k-means (b), pre $k = 4 \ldots 30$ pri aplikácií na 6. periódu.



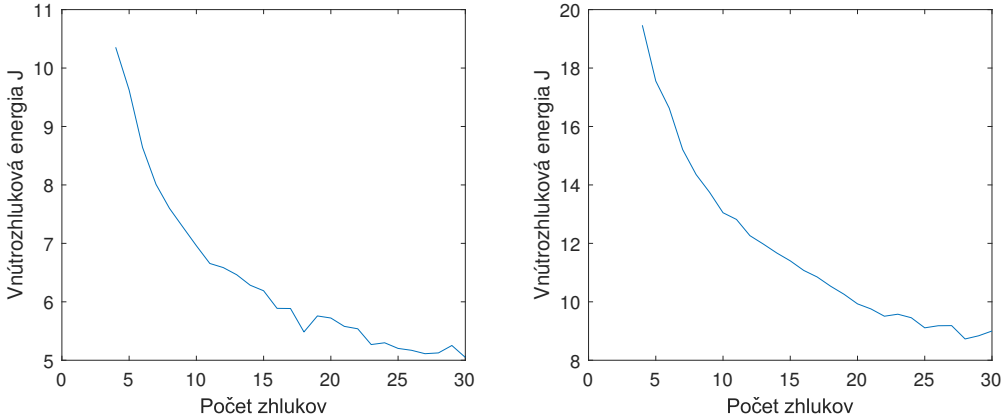
(a) k-modes (b) k-means

Obr. 4.8: Priebeh obrysových koeficientov pre metódu k-modes (a), metódu k-means (b), pre $k = 4 \ldots 30$ pri aplikácií na 6. periódu.



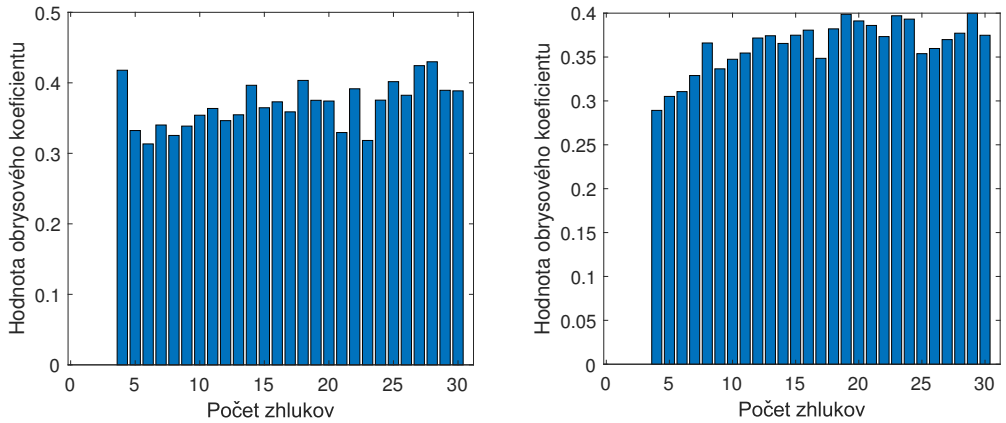
(a) k-modes (b) k-means

Obr. 4.9: Riešenie zhlukovania 6. periódy pre k-modes, $k = 20$ (a) s hodnotou obrysového koeficientu $\overline{s}(k) = 0,42$ a pre k-means, $k = 26$ (b) s hodnotou obrysového koeficientu $\overline{s}(k) = 0,42$.



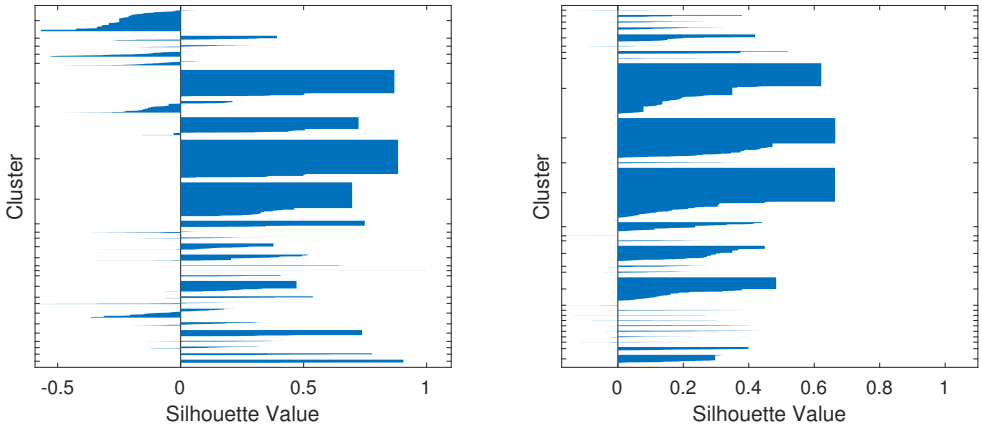
(a) k-modes (b) k-means

Obr. 4.10: Priebeh vnútrozhlukovej energie J pre metódu k-modes (a), metódu k-means (b), pre $k = 4 \ldots 30$ pri aplikácií na obe periódy.



(a) k-modes (b) k-means

Obr. 4.11: Priebeh obrysových koeficientov pre metódu k-modes (a), metódu k-means (b), pre $k = 4 \ldots 30$ pri aplikácií na obe periódy.



(a) k-modes (b) k-means

Obr. 4.12: Riešenie zhlukovania oboch období pre k-modes, $k = 28$ (a) s hodnotou obrysového koeficientu $\overline{s}(k) = 0,43$ a pre k-means, $k = 29$ (b) s hodnotou obrysového koeficientu $\overline{s}(k) = 0,40$.

daného zhluku. Ďalej si môžeme všimnúť, že viac než 70% jedincov je koncentrovaných v ôsmich mnohopočetných zhlukoch. Pre zhluk číslo 1 dokonca neexistuje titul s výrazným zastúpením akejkoľvek kategórie, väčšina zastúpenia sa pohybuje do 10%, čo značí, že pre tento zhluk bolo charakteristické kategorizované tituly skôr nemať, ako mať.

Obdobnú tabuľku 4.3 sme vytvorili pre výsledky algoritmu k-means, $k = 29$, zobrazené na 4.12b, kde je vidieť podobné správanie zhlukov ako v pre predchádzajúcu metódu. Z tohto dôvodu sme vytvorili tabuľku 4.4, ktorá mapuje podobnosť medzi veľkými zhlukmi metódy k-means so zhlukmi metódy k-modes. Tabuľka obsahuje index (číslo) zhluku totožné s predchádzajúcimi tabuľkami, veľkosť zhluku pre obe metódy, vzdialenosť centier nájdených rôznymi metódami a počet prvkov, ktoré sa zhodujú v zaradení oboch algoritmov. Možno pozorovať, že takmer všetky objekty konkrétneho zhluky v tabuľke určené metódou k-modes sa vyskytujú v jednom zhluku uvedenom v rovnakom riadku pre metódu k-means. Môžeme si povšimnúť, že tieto dvojice zhlukov disponujú takmer rovnakými významnými kategóriami (výnimkou je dvojica 10-10, kde metóda k-modes ukazuje pre zhluk 10 kategórie 6 a 7, zatiaľ čo metóda k-means tieto kategórie zúžila len na kategóriu 7). Tabuľka vzdialeností jednotlivých centier 4.21 sa nachádza v prílohe 1.

Poslednou tabuľkou uvedenou v tejto časti je tabuľka 4.5, kde sa podobným princípom snažíme prepojiť na seba zhluky, ktoré vznikli zhlukovaním metódou k-modes zvlášť v 5. perióde znázornené na 4.6a a zvlášť v 6. perióde s vizualizáciou 4.9a. Túto voľbu sme učinili na základe toho, že metóda k-modes mala vyššiu hodnotu obrysového koeficientu pre každú z období, v porovnaní s metódou k-means. Do tabuľky sme zaznačili pre periódy zvlášť číslo zhluku, počet objektov pre daný zhluk, priemernú a maximálnu vzdialenosť jedinca od módu a charakteristickú kategóriu. Do tabuľky sme umiestnili také zhluky, ktorých centrá sú plne totožné, pričom charakteristické kategórie sú takmer nezmenené (výnimku tvorí dvojica 5-3). Je vidieť, že v 6. perióde je počet objektov v zhlukoch menší než pre ich ekvivalent z 5. periódy, pripomeňme, že dataset 5. periódy má 2343 záznamov s aspoň jedným titulom a dataset 6. periódy sa zakladá z 1201 objektov s minimálne jednou kategóriou titulu. Tabuľka vzdialeností jednotlivých centier 4.20 sa nachádza v prílohe 1.

V priebehu tejto sekcie o základnom datasete sme sa stretli s tým, že pre našu egyptologickú úlohu neboli metódy určenia počtu zhlukov v kombinácií algoritmami vo svojich základných podobách dostatočne jasné. Napriek tomu sa nám na tejto diskrétnej úlohe podarilo demonštrovať,

že obe metódy boli schopné detekovať podobne veľké a výrazné zhluky. Nakoniec sme v dátach spozorovali, že niektoré zhluky pretrvávajú naprieč periódami 5 a 6.

$\mathcal{C}$	N	d_{rq}	$\overline{r}$	r_{max}	najzastúpenejšie kategórie
1	310	1	1,9	10	
2	70	1	1,3	13	2 3 8
3	30	9	8,7	15	2 4 5 17 22 29 41
4	72	2	3,9	10	2 5
5	60	3	5,5	9	1 4 7 24
6	387	1	0,2	4	8
7	170	2	2,1	9	2
8	263	1	0,4	6	5
9	557	1	0,1	4	1
10	494	1	0,4	5	6 7
11	91	1	0,3	3	1 8
12	20	5	7,5	12	2 4 5 7 15
13	20	8	6,5	11	1 2 5 7 15 20 41
14	96	1	1,4	7	2 3
15	94	3	1,5	6	5 7 20
16	7	1	0,6	2	2 8 22
17	1	7	0,0	0	1 2 4 6 7 8 9 25 58 60
18	16	1	0,9	4	2 22
19	158	1	0,6	5	5 8
20	27	1	0,6	4	6 7 8
21	20	2	4,0	8	1 2 21
22	131	3	2,5	8	4 9
23	48	3	4,3	9	2 3 4 5 7 12
24	90	1	0,4	5	1 6 7
25	17	9	5,4	13	4 7 9 11 12 29
26	32	3	2,8	7	4 9 47
27	26	2	0,6	5	2 28
28	46	2	0,2	3	16

Tabuľka 4.2: Záznam výsledkov zhlukovania záznamov oboch období pre k-modes, $k = 28$ znázorneného na 4.12a. Tabuľka obsahuje číslo zhluku $\mathcal{C}$, počet jedincov v zhluku N, vzdialenosť d_{rq} centra zhluku $\mathcal{C}_r$ od cudzieho najbližšieho zhluku $\mathcal{C}_q$, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhluku disponovalo danou kategóriou).

$\mathcal{C}$	N	d_{rq}	$\overline{r}$	r_{max}	najzastúpenejšie kategórie
1	7	8,9	10,2	11,9	2 5 7 22 29 33 41
2	29	3,6	3,0	8,8	11
3	21	5,5	6,5	9,1	2 4 5 7 9 24
4	37	5,8	6,6	9,4	2 3 5 7 12
5	104	2,7	2,3	5,3	
6	8	9,8	11,5	12,5	1 2 4 5 7 9 12 15 17 22 29 33 36 47
7	41	2,0	1,7	6,7	6 7 10
8	6	6,4	4,3	6,3	7 8 24
9	746	2,4	1,7	6,8	8
10	588	1,9	1,3	7,0	7
11	15	7,7	6,9	8,8	1 4 5 12 29 44
12	738	2,4	1,6	8,1	1
13	135	3,9	3,5	9,6	5 7 20
14	2	12,1	8,5	8,5	1 2 5 7 13 15 17 20 24 27 30 31 35 51
15	12	8,5	9,0	10,1	1 2 3 4 5 7 15 30 31
16	226	3,3	2,8	8,2	2
17	23	5,8	6,7	9,7	2 4 7 9 20 27
18	18	5,5	6,2	9,2	4 5 7 9 24
19	344	2,5	2,3	10,2	5
20	5	9,9	9,8	11,4	1 2 4 5 7 9 24 30 31
21	11	8,5	7,8	9,1	2 3 4 5
22	7	10,4	8,8	12,1	1 2 4 5 7 9 12 15 17 24 29 33 41
23	9	7,7	7,2	9,0	1 2 4 7 12 20 29
24	13	7,5	6,6	9,7	4 7 9 11 12 29
25	17	7,7	7,1	10,1	1 2 3 5 7 15
26	15	7,1	7,7	10,3	2 4 5 9 47
27	8	6,4	5,5	11,1	5 7 8 24
28	50	2,7	2,3	5,1	14
29	118	1,9	2,1	8,8	1 6 7

Tabuľka 4.3: Záznam výsledkov zhlukovania záznamov oboch periód pre k-means, $k = 29$ znázorneného na 4.12b. Tabuľka obsahuje číslo zhluku $\mathcal{C}$, počet jedincov v zhluku N, vzdialenosť d_{rq} centra zhluku $\mathcal{C}_r$ od cudzieho najbližšieho zhluku $\mathcal{C}_q$, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhluku disponovalo danou kategóriou).

$\mathcal{C}_{\text{means}}$	N_{means}	$\mathcal{C}_{\text{modes}}$	N_{modes}	d_{rq}	$\text{N}_{\text{zhodné}}$
9	746	6	387	1,0	384
10	588	10	494	0,7	459
12	738	9	557	0,9	555
19	344	8	263	1,3	227

Tabuľka 4.4: Mapovanie najpočetnejších zhlukov riešenia k-means na najpodobnejšie riešenie k-modes. Najpodobnejšie v zmysle najmenšej vzdialenosti ich centroidov k módom. Riešenie bolo pre dáta z oboch periód. Tabuľka obsahuje číslo zhluku pre k-means $\mathcal{C}_{\text{means}}$, počet jedincov pre k-means N_{means}, číslo zhluku pre k-modes $\mathcal{C}_{\text{modes}}$, počet jedincov pre k-modes N_{modes}, vzdialenosť centra a módu d_{rq}, a počet jedincov, ktorí sú zhodne zaradení do oboch týchto zhlukov $\text{N}_{\text{zhodné}}$.

5. perióda					6. perióda				
$\mathcal{C}$	N	$\overline{r}$	r_{max}	kategórie	$\mathcal{C}$	N	$\overline{r}$	r_{max}	kategórie
1	499	0,6	5	6 7	4	206	0,8	4	6 7
5	133	1,9	6		3	100	2,1	7	2
7	51	0,2	3	1 8	20	41	0,3	4	1 8
9	72	1,9	5	4	5	58	1,7	6	4
13	258	0,2	6	8	12	141	0,1	2	8
14	187	0,4	3	5	8	72	0,3	3	5
19	102	0,7	3	5 8	17	75	0,6	4	5 8
23	33	0,0	1	16	7	19	0,1	1	16
26	63	1,0	4	2 3	16	49	1,7	7	2 3

Tabuľka 4.5: Mapovanie zhlukov 5. periódy na zhluky 6. periódy pre metódu k-modes. Tabuľka obsahuje pre každú periódu číslo zhluku $\mathcal{C}$, počet jedincov v danom zhluku N, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhlukov disponovalo danou kategóriou).

4.2 Rozšírený dataset

Na rozšírenom datasete o kód hrobky, typ hrobky, záznam o skutočnosti, kto bol majiteľom, resp. pochovaným v danej hrobke, alebo len zmienkou zostavíme úlohu zhlukovania hrobiek na základe titulov majiteľov a typu hrobky respektíve na základe všetkých titulov zaznamenaných v hrobke a jej type. Budeme využívať zaradenie titulov do kategórií titulov, ktoré sú uvedené na histograme 4.3 a typy hrobiek, ktoré sme uviedli na histograme4.2.

4.2.1 Predspracovanie dát

Rozšírený dataset sme obdržali vo forme niekoľkých *.csv* tabuliek, schematicky znázornené na 4.1. Nami implementované klastrovacie funkcie v prostredí MATLAB využívajú výlučne číselnú vstupnú maticu, preto sme implementovali niekoľko vlastných funkcií umožňujúcich prevod tabuliek do požadovanej matice. Najprv sme využili možnosť importovať tabuľku do matice reťazcov (string). Pomocou vlastnej funkcie sme uskutočnili selekciu potrebných stĺpcov (ID osoby, Kód hrobky a neskôr po aktualizácií databázy údaj, či šlo o vlastníka hrobky). Ďalej sme vytvorili funkciu, ktorá prepojí dve rôzne tabuľky na základe zhody v zvolených stĺpcoch a mohli sme tak pripojiť typ hrobky a prepojiť titul s kategóriou. V ďalšom kroku sme zoskupili niekoľko podobných kategórí do jednej, viď úvod kapitoly 4 a histogram 4.3. Funkciou vyberajúcou unikátne riadky v reťazcovej matici sme sa zbavili duplikátov, čo nám následne umožnilo premapovať kategórie a typy hrobiek na binárny priestor a zostaviť odpovedajúcu maticu reťazcov, ktorú prevedieme do numerickej matice.

Uvedeným spôsobom si pripravíme dva typy vstupných dátových matíc do zhlukovacích úloh. Maticu $\mathbf{X}$, ktorej objekty (riadky) predstavujú konkrétne hrobky vytvoríme binárne tak, že stĺpce odpovedajú jednotlivým kategóriám titulov, čomu zodpovedá hodnota 1 ak titul bol uvedený v hrobke a 0 ak nebol, ďalej stĺpce odpovedajúce tomu, akého typu bola hrobka pričom 1 ak bola daného typu a 0 ak nebola. Nulové stĺpce odstránime. Pre presnú predstavu viď prílohu 4.14, 4.15. Maticu $\mathbf{Y}$ zostavíme obdobne, pridáme stĺpec ktorý ponesie informáciu o počte ľudí zaznamenaných v hrobke, respektíve pochovaných v hrobke.

Úlohu doplníme vytvorením matice $\mathbf{Z}$, ktoré nebude obsahovať typ hrobky. Chceme pre konkrétne k posúdiť, či typ hrobky hrá v posudzovaní zhlukov dôležitú úlohu.

X	Kategória 1	⋯	Kategória 26	Typ hrobky 1	⋯	Typ hrobky 22
	$\{0;1\}$	⋯	$\{0;1\}$	$\{0;1\}$	⋯	$\{0;1\}$

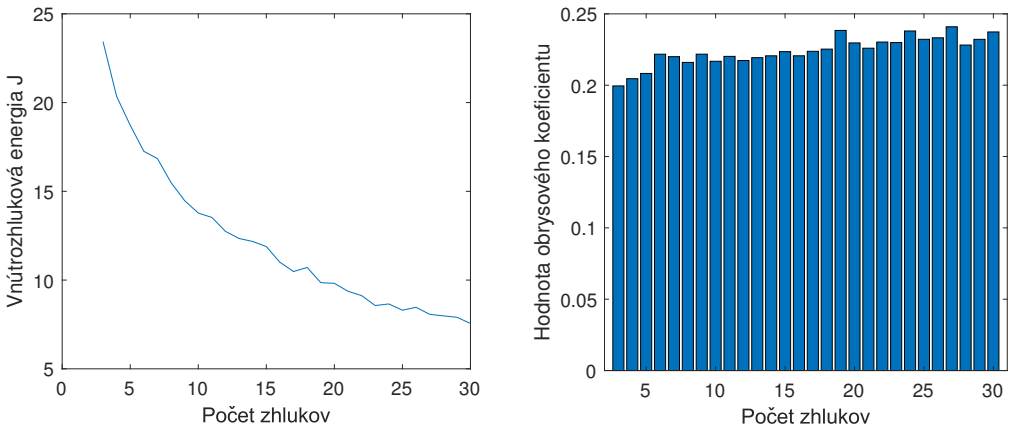
Y	Počet ľudí	Kategória 1	⋯	Kategória 26	Typ hrobky 1	⋯	Typ hrobky 22
		$\{0;1\}$	⋯	$\{0;1\}$	$\{0;1\}$	⋯	$\{0;1\}$

Z	Kategória 1	Kategória 2	⋯	Kategória 26
	$\{0;1\}$	$\{0;1\}$	⋯	$\{0;1\}$

4.2.2 Aplikácia zhlukovacích metód

Za zhlukovaciu metódu sme zvolili metódu k-means (viď algoritmus 3) s blokovou metrikou (1.2). V predchádzajúcej časti so základným datasetom 4.1 sme pozorovali podobné správanie medzi k-modes a k-means s blokovou metrikou. Z dôvodu použitia príznaku počtu pochovaných (zmienených) ako kvantitatívnej premennej nemožno využiť metódu módov na dataset $\mathbf{Y}$, preto sa uspokojíme s metódou k-means a konzistentne ju aplikujeme aj na dataset $\mathbf{X}$. Poznamenajme, že ide o rovnakú pozmenenú implementáciu ako v úlohe na základnom datasete v časti 4.1.2.

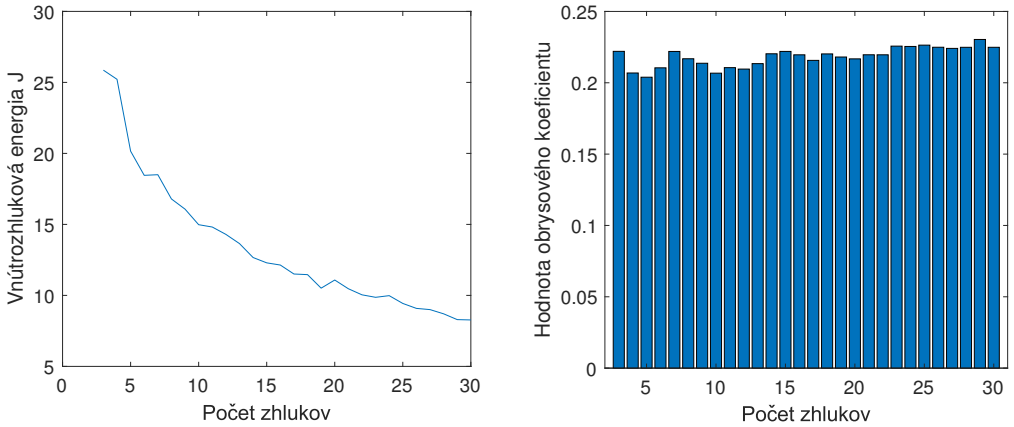
Postupujeme tak, že z vstupných dát $\mathbf{X}$ vyberieme podmnožinu hrobiek s vlastnosťami zdedenými výlučne po pochovaných osobách (majiteľoch hrobky), označíme $\mathbf{X}^\dagger$ a podmnožinu celku, ktorú chápeme ako hrobky s vlastnosťami po všetkých osobách, ktoré sú v nej zmienené $\mathbf{X}$, rovnako postupujeme pre $\mathbf{Y}$ a vyberáme pri rovnakom značení $\mathbf{Y}^\dagger$,$\mathbf{Y}$. Na tieto podmnožiny použijeme metódu k-means pre rôzne počty zhlukov k ($k = 3, \ldots, 30$) v sto jedenástich opakovaniach, pričom za ponechané riešenie považujeme také, ktorému sa zlepší obrysový koeficient. Vykreslíme priebehy vnútrozhlukových energií J a graf obrysových koeficientov, pričom sa podľa popisu v časti 3.1 metódy kolena a obrysového koeficientu zvoliť vhodné k.



(a) (b)

Obr. 4.13: Priebeh vnútrozhlukovej energie J a obrysových koeficientov k-means zhlukovania datasetu hrobiek $\mathbf{X}^\dagger$ podľa titulov vlastníkov a typu hrobky bez príznaku počtu pochovaných v danej hrobke, pre $k = 3 \cdots k = 30$

Po vykreslení priebehu vnútrozhlukovej energie J grafom 4.13a a obrysového grafu 4.13b pre dataset $\mathbf{X}^\dagger$ možno konštatovať, že jednoznačne rozlíšiteľné koleno sa na grafe energie J nenachádza rovnako ani značne vyššia hodnota obrysového koeficientu. Najvyššiu obrysový koeficient hodnotu dosahuje pre $k = 27$ a to $\overline{s}(k) = 0,241$. Poznamenajme, že pre $k = 19$ je $\overline{s}(k) = 0,238$ a pre $k = 7$ dosahuje $\overline{s}(k) = 0,220$.

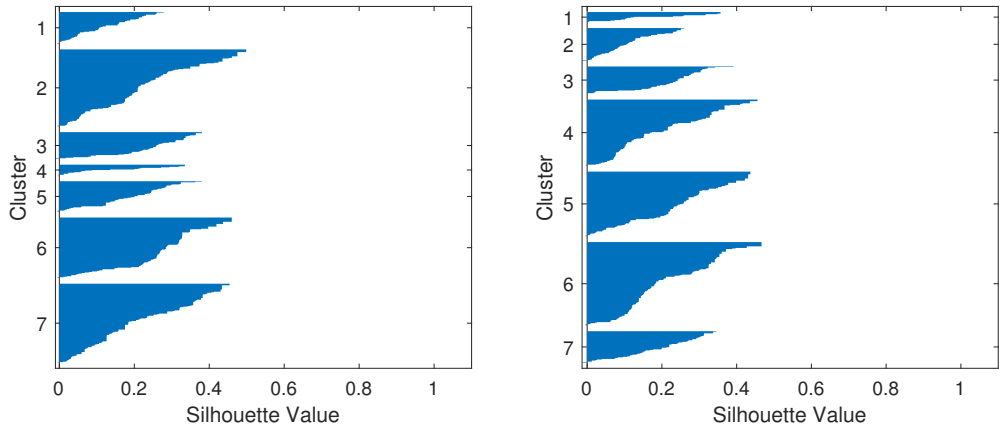


(a) (b)

Obr. 4.14: Priebeh vnútrozhlukovej energie J a obrysových koeficientov k-means zhlukovania datasetu $\mathbf{Y}^{\dagger}$ hrobiek podľa titulov vlastníkov a typu hrobky s príznakom počtu pochovaných v danej hrobke, pre $k = 3 \cdots k = 30$

Podobná situácia nastáva aj pre dataset $\mathbf{Y}^{\dagger}$, v ktorom berieme do úvahy počet pochovaných, ktorý nadobúda hodnoty 1,2 alebo 3, kedy na grafe J 4.14a nevieme rozpoznať jednoznačné koleno a obrysový graf 4.14b nezobrazuje markantne najvyššiu hodnotu obrysového koeficientu, ktorá nastáva pre $k = 29$ a má hodnotu $\overline{s}(k) = 0,230$. Podobne na tom je zhlukovanie pre $k = 3$ a $k = 7$, kedy dosahujeme $\overline{s}(k) = 0,222$.

Keďže sa výsledné hodnoty obrysových koeficientov priveľmi nelíšia, rozhodli sme sa pre uľahčenie posudzovania výsledkov prijať riešenia pre $k = 7$ ako pre $\mathbf{X}^{\dagger}$, tak aj pre $\mathbf{Y}^{\dagger}$. Toto riešenie znázorníme na obrázku 4.15, kde pozorujeme, že hoci hodnoty obrysov nedosahujú vysoké hodnoty, tak takmer nepresahujú do zápornej časti a pre oba prípady majú zhluky podobné tvary a šírky.

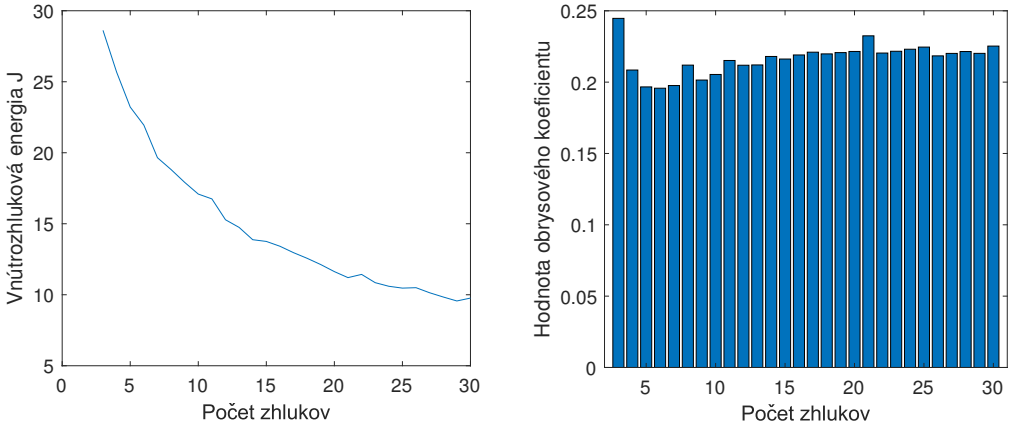


(a) $\mathbf{X}^{\dagger}$ bez príznaku počtu majiteľov (b) $\mathbf{Y}^{\dagger}$ s príznakom počtu majiteľov

Obr. 4.15: Obrysový graf výsledných zhlukov k-means hrobiek s redukciou na majiteľov (pochovaných), titulov a typov hrobiek pre zvolené $k = 7$

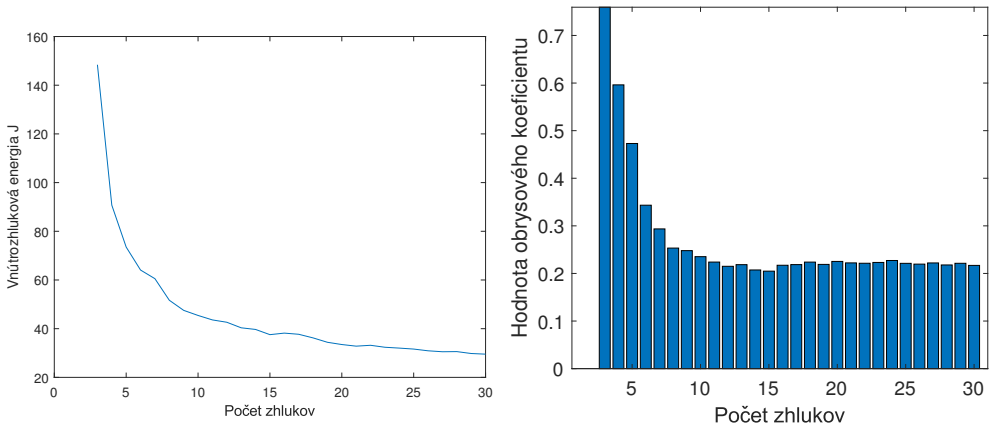
Pre dataset $\mathbf{X}$ opäť na grafe energie J 4.16a nepozorujeme koleno a na obrysovom grafe pozorujeme najvyššiu hodnotu obrysového koeficientu pre $k = 3$ s hodnotou $\overline{s}(k) = 0.245$, pričom pre $k = 8$ máme $\overline{s}(k) = 0,212$.

V poslednom datasete **Y** udávame počty zmienok o ľuďoch, pričom máme hrobky, kde sa nachádza napríklad 323, 29 či napríklad 1 zmienka, celkom máme 46 rôznych počtov zmienok, čo kladie vysokú váhu tomuto príznaku. Všimneme si, že na grafe J 4.17a pozorujeme koleno v $k = 8$, avšak najvyššia hodnota obrysového koeficientu podľa obr.: 4.17b leží v $k = 3$ a dosahuje $\overline{s}(k) = 0,759$, pričom napríklad v kolene $k = 8$ má hodnotu $\overline{s}(k) = 0,253$.



(a) (b)

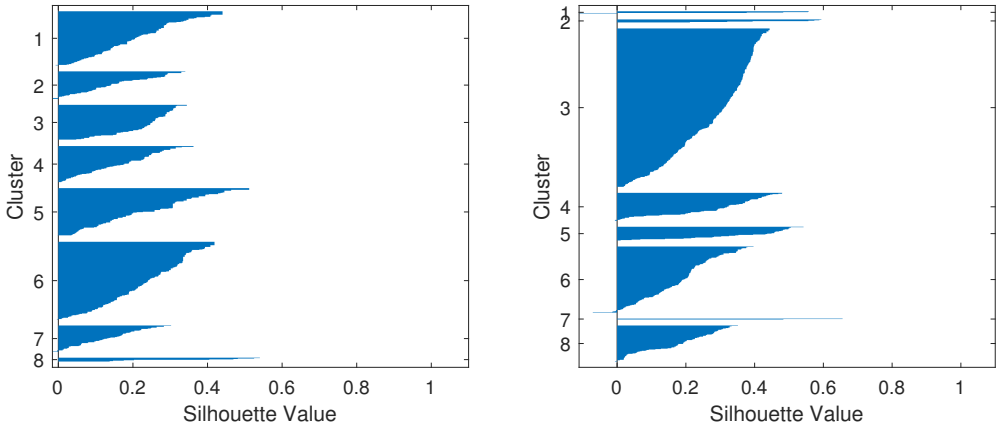
Obr. 4.16: Priebeh vnútrozhlukovej energie J a obrysových koeficientov k-means zhlukovania datasetu hrobiek **X** podľa titulov v nich zaznamenaných bez príznaku počtu zmienených ľudí v danej hrobke a typu hrobku pre $k = 3 \cdots k = 30$



(a) (b)

Obr. 4.17: Priebeh vnútrozhlukovej energie J a obrysových koeficientov k-means zhlukovania datasetu hrobiek **Y** podľa titulov v nich zaznamenaných s príznakom počtu zmienených ľudí v danej hrobke a typu hrobky pre $k = 3 \cdots k = 30$

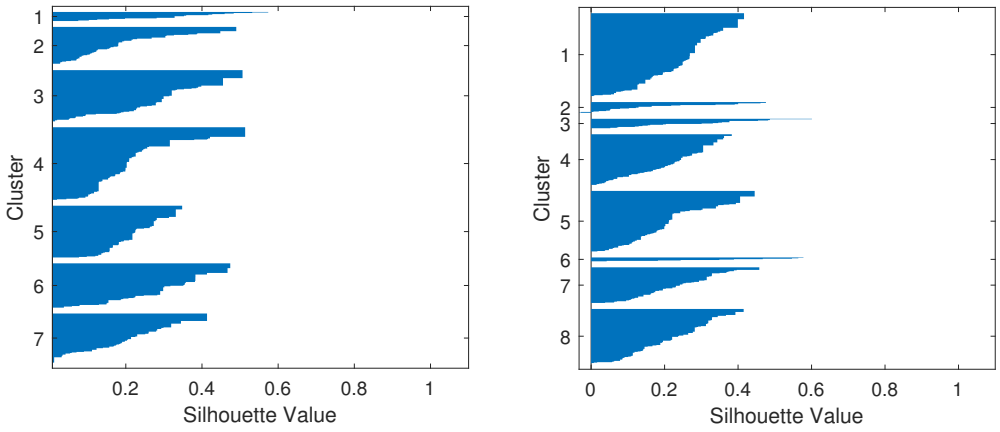
Podľa metódy kolena prijímame $k = 8$ pre oba datasety **X** aj **Y**, riešenie sme vykreslili na grafe 4.18. Pozorujeme, že siluety vzniknutých zhlukov strádajú na podobnosti. Všimneme si, že pre **Y** sa objavujú 3 nízkopočetné zhluky a 1 vysokopočetný, čo je zapríčinené nevyrovnanosťou v počte zmienok. Ak porovnávame zmienku 1 človeka, 29 ľudí alebo 323 ľudí, oproti zvyšným binárnym príznakom sa v dimenzii počtu dostávame nepomerne priďaleko, čo má za následok práve vylúčenie týchto zhlukov na periférium, kam už binárne podobný ostatok nedosiahne.



(a) **X** bez príznaku počtu zmienok ľudí (b) **Y**s príznakom počtu zmienok ľudí

Obr. 4.18: Obrysový graf výsledných zhlukov k-means hrobiek pre všetky zaznamenané osoby, ich tituly a typ hrobky pre zvolené $k = 8$

Posledným krokom je aplikovať k-means na datasety $\mathbf{Z}^{\dagger}$ (pochovaných) pre $k = 7$ a $\mathbf{Z}$ (všetkých zmienok) pre $k = 8$, aby sme mohli posúdiť relevantnosť vplyvu typu hrobky na zhlukovanie.



(a) $\mathbf{Z}^{\dagger}$ bez príznaku počtu pochovaných ľudí, bez typu hrobky,$k = 7$ (b) **Z** bez príznaku počtu zmienok ľudí, bez typu hrobky, $k = 8$

Obr. 4.19: Obrysový graf výsledných zhlukov k-means hrobiek podľa ich titulov pre zvolené $k = 7$ a $k = 8$

4.2.3 Vyhodnotenie

V tomto oddieli porovnáme výsledky zhlukovania pre našu úlohu pre jednotlivé subdatasety.

V tabuľke 4.6 uvádzame prehľad 7 zhlukov vzniknutých aplikáciou k-means, $k = 7$ na podmnožinu hrobiek podľa vlastností majiteľov a typu hrobky, obr.:4.15a. V tabuľke uvádzame pre konkrétny zhluk $\mathcal{C}_r$, počet jedincov v zhluku N, vzdialenosť d_{rq} centra zhluku $\mathcal{C}_r$ od cudzieho najbližšieho zhluku $\mathcal{C}_q$, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhluku disponovalo danou kategóriou). Na základe tabuľky z prílohy 4.15 teda môžeme povedať, že pre hrobky zhluku 1 nemáme ani dominantný titul, ani dominantný typ hrobky, zhluk 2 je charakteristický titulmi z kategórií 1 *administrative title* a

9 *epiteth*, zhluku 3 dominuje typ hrobky 43 *unknown*, zhluku 6 typ hrobky 32 *rock-cut tomb*. Všimneme si že vzdialenosť akéhokoľvek zhluku od iného jemu najbližšieho cudzieho zhluku je menšia, než priemerná vzdialenosť vlastných objektov od svojho centra.

Na podobnom princípe popíšeme tabuľky 4.7, 4.8, 4.9.

$\mathcal{C}$	N	d_{rq}	$\overline{r}$	r_{max}	najzastúpenejšie kategórie
1	59	2,7	4,4	6,6	
2	144	3,35	3,8	6,4	1 9
3	50	2,9	3,7	5,5	43
4	20	4,6	4,7	6,3	9 18 19
5	57	4,5	4,7	7,4	1 12 18 19 21 24
6	113	2,7	3,6	7,1	32
7	148	3,3	3,9	5,9	

Tabuľka 4.6: Záznam výsledkov zhlukovania hrobiek datasetu $\mathbf{X}^{\dagger}$, metódou k-means pre $k = 7$ znázorneného na obrázku 4.15a. Tabuľka obsahuje číslo zhluku $\mathcal{C}$, počet jedincov v zhluku N, vzdialenosť d_{rq} centra zhluku $\mathcal{C}_r$ od cudzieho najbližšieho zhluku $\mathcal{C}_q$, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhluku disponovalo danou kategóriou).

V tabuľke 4.7 datasetu majiteľov $\mathbf{Z}^{\dagger}$ 4.19a hrobiek pre redukovaný príznakový priestor len na kategórie titulov vidíme, že v zhlukoch 1-4 dominujú *unlisted* tituly (24) a *rank title* 21. Okrem zhluku 1 a 2 opäť platí, že centrá sú bližšie iným centrám, než priemerne vlastným prvkom. O zhluku 1 navyše môžeme povedať, že od najbližšieho centra je ďalej než od svojho najvzdialenejšieho jednica, naviac ide o nízkopočetný zhluk.

$\mathcal{C}$	N	d_{rq}	$\overline{r}$	r_{max}	najzastúpenejšie kategórie
1	17	4,1	3,0	3,9	1 12 18 21 24
2	70	3,1	3,0	4,9	1 9 18 19 20 21 24
3	96	2,0	2,3	4,7	24
4	136	2,2	2,5	5,6	9 21 24
5	97	2,0	2,5	5,0	
6	83	2,1	2,3	3,7	1
7	92	2,2	2,7	4,6	9

Tabuľka 4.7: Záznam výsledkov zhlukovania hrobiek datasetu $\mathbf{Z}^{\dagger}$, metódou k-means pre $k = 7$ znázorneného na obrázku 4.19a. Tabuľka obsahuje číslo zhluku $\mathcal{C}$, počet jedincov v zhluku N, vzdialenosť d_{rq} centra zhluku $\mathcal{C}_r$ od cudzieho najbližšieho zhluku $\mathcal{C}_q$, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhluku disponovalo danou kategóriou).

Riešenie vykreslené grafom 4.18a k-means, $k = 8$ pre dataset hrobiek podľa všetkých uvedených zmienok titulov a typu hrobky $\mathbf{X}$ uvádzame v tabuľke 4.8. Všimneme si, že vzniká pomerne dobre separovateľný malý zhluk s číslom 8 pre deväť hrobiek, ktoré sú charakteristické množstvom titulov, ale nie typom hrobky. Pre stodvadsaťšesť hrobiek v zhluku 1 je typické byť typu *rock-cut tobm* (35). Majoritnnými titulmi zhluku 2 sú napríklad tituly z kategótie 20 *privacy of King*. Vidíme že zhluky spájajú kategórie 9, 19 a 22 *epithet, priestly title* a *rank title*. A vidíme, že v zhlukoch 7 a 8 sa vyskytuje majoritne *family relation* (10).

Poslednou tabuľkou tohto typu je 4.9, ktorá znázorňuje riešenie 4.19b pre dataset $\mathbf{Z}$ hrobiek všetkých kategórií titulov na zúženom príznakovom priestore bez typu hrobky. Môžeme si

všimnúť, že dostávame veľmi podobný výčet charakteristických kategórií ako pre riešenie s uvažovaním typu hrobky 4.8.

Na záver ukazujeme mapovanie hrobiek s uvažovaním titulov výhradne majiteľov resp. s uvažovaním každého zaznamenaného titulu z redukovaného príznakového priestoru na rozšírený priestor o typ hrobky: teda z $\mathbf{Z}^\dagger$ na $\mathbf{X}^\dagger$ resp.z $\mathbf{Z}$ na $\mathbf{X}$. A mapovanie hrobiek s uvažovaním titulov výhradne majiteľov medzi priestoro bez príznaku počtu pochovaných na priestor s týmto príznakom $\mathbf{X}^\dagger$ na $\mathbf{Y}^\dagger$.

V tabuľke 4.10 vidíme, že jedine malý zhluk číslo 1 sa celý premapoval do zhluku 5 v druhom priestore, kde však predstavuje minoritnú časť prvkov. Pre ostatné zhluky sa súdržne premapovala približne polovica objektov, pričom napríklad zhluk 2 z redukovaného priestoru tvorí základ zhluku 5 v rozšírenom príznakovom priestore (zhluk 2 z tabuľky 4.7 a zhluk 5 z tabuľky 4.6 majú spoločné charakteristické príznaky pre kategórie 1,18,19,21 a 24). V tomto prípade redukcia príznakového priestoru mala za následok iné zhlukovacie riešenie.

Pre zhlukovanie hrobiek podľa všetkých zmienok pozorujeme v prestupovej tabuľke 4.11, že zachovanie zhlukov je minimálne. Súdržný zostáva zhluk 8, ktorý sa celý premapije do zhluku 8 v rozšírenom priestote, kde však netvorí ani polovicu obsahu. Žiadne iné výrazné premapovanie nenastáva. Napriek relatívnej podobnosti obrysov grafov 4.15a a 4.19a, 4.18a a 4.19b sa

$\mathcal{C}$	N	d_{rq}	$\overline{r}$	r_{max}	najzastúpenejšie kategórie
1	126	3,5	4,1	7,3	35
2	64	4,6	4,9	7,4	1 13 19 20 22 26
3	82	3,0	3,9	6,6	47
4	84	3,0	4,3	5,8	
5	111	3,7	3,9	7,0	
6	181	3,9	3,9	6,9	9 19 22 41
7	61	4,0	4,9	6,9	1 9 10 19 22 26
8	9	8,1	5,1	6,1	1 3 9 10 13 14 19 20 21 22 24 26

Tabuľka 4.8: Záznam výsledkov zhlukovania hrobiek datasetu $\mathbf{X}$, metódou k-means pre $k =$ znázorneného na obrázku 4.18a. Tabuľka obsahuje číslo zhluku $\mathcal{C}$, počet jedincov v zhluku N, vzdialenosť d_{rq} centra zhluku $\mathcal{C}_r$ od cudzieho najbližšieho zhluku $\mathcal{C}_q$, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhluku disponovalo danou kategóriou).

$\mathcal{C}$	N	d_{rq}	$\overline{r}$	r_{max}	najzastúpenejšie kategórie
1	193	2,0	2,7	5,1	
2	25	3,9	3,4	4,6	9 20 21
3	22	4,9	3,5	4,7	1 13 19 22 26
4	118	2,4	2,9	5,7	1 9 19 21 22
5	141	2,4	2,7	5,3	9 19 22 26
6	9	6,9	3,9	4,7	1 3 9 10 13 14 19 20 21 22 24 26
7	83	2,6	3,0	5,4	1 9 19 20 21 22 26
8	127	2,0	2,8	5,6	1

Tabuľka 4.9: Záznam výsledkov zhlukovania hrobiek datasetu $\mathbf{Z}$, metódou k-means pre $k =$ znázorneného na obrázku 4.19b. Tabuľka obsahuje číslo zhluku $\mathcal{C}$, počet jedincov v zhluku N, vzdialenosť d_{rq} centra zhluku $\mathcal{C}_r$ od cudzieho najbližšieho zhluku $\mathcal{C}_q$, priemernú vzdialenosť objektov od svojho centra $\overline{r}$, maximálnu vzdialenosť objektu od svojho centra r_{max} a najzastúpenejšie kategórie titulov charakterizujúce daný zhluk (aspoň 85% objektov zhluku disponovalo danou kategóriou).

		$\mathbf{X}^\dagger$							
		1	2	3	4	5	6	7	$N_{Z^\dagger}$
$\mathbf{Z}^\dagger$	1	0	0	0	0	17	0	0	17
	2	0	23	1	1	40	5	0	70
	3	6	2	11	2	0	24	51	96
	4	24	55	8	1	0	31	17	136
	5	5	0	12	4	0	32	44	97
	6	18	4	9	0	0	16	36	83
	7	6	60	9	12	0	5	0	92
	$N_{X^\dagger}$	59	144	50	20	57	113	148	

Tabuľka 4.10: Mapovanie zhlukov datasetu $\mathbf{Z}^\dagger$ na zhluky $\mathbf{X}^\dagger$. Riadok tabuľky uvádza číslo zhluku výsledku k-means, $k = 7$ na $\mathbf{Z}^\dagger$ a postupne udáva koľko jeho vlastných jedincov sa vyskytovalo v zhlukoch 1 až 7 k-means, $k = 7$ datasetu $\mathbf{X}^\dagger$. Posledný stĺpec udáva počet jedincov v danom zhluku pre $\mathbf{Z}^\dagger$ a posledný riadok naopak počet jedincov v danom zhluku pre $\mathbf{X}^\dagger$.

podobnosť v zostavení zhlukov nepotvrdila.

		$\mathbf{X}$								
		1	2	3	4	5	6	7	8	$N_{\mathbf{Z}}$
$\mathbf{Z}$	1	53	1	19	8	2	42	1	0	126
	2	0	18	0	3	13	0	20	10	64
	3	5	1	22	13	3	37	1	0	82
	4	17	0	25	9	2	31	0	0	84
	5	2	0	27	0	0	82	0	0	111
	6	59	0	20	60	15	2	25	0	181
	7	6	0	0	23	3	0	28	1	61
	8	0	0	0	0	0	0	0	9	9
	$N_{\mathbf{X}}$	142	20	113	116	38	194	75	20	

Tabuľka 4.11: Mapovanie zhlukov datasetu $\mathbf{Z}$ na zhluky $\mathbf{X}$. Riadok tabuľky uvádza číslo zhluku výsledku k-means, $k = 8$ na $\mathbf{Z}$ a postupne udáva koľko jeho vlastných jedincov sa vyskytovalo v zhlukoch 1 až 8 k-means, $k = 8$ datasetu $\mathbf{X}$. Posledný stĺpec udáva počet jedincov v danom zhluku pre $\mathbf{Z}$ a posledný riadok naopak počet jedincov v danom zhluku pre $\mathbf{X}$.

Posledným krokom je prepojenie priestoru hrobiek podľa majiteľov a typov hrobiek $\mathbf{X}^\dagger$ na priestor rozšírený o počet pochovaných $\mathbf{Y}^\dagger$, uvádzame v tabuľke 4.12. V tomto prípade pozorujeme trend výrazného sa zachovania zloženia zhluku, čo podporuje úvahu, že počet pochovaných výrazne neovplyvnil výsledok zhlukovania.

		$\mathbf{Y}^\dagger$							
		1	2	3	4	5	6	7	$N_{\mathbf{X}^\dagger}$
$\mathbf{X}^\dagger$	1	0	1	0	18	0	0	0	19
	2	28	23	1	0	0	9	0	61
	3	2	0	48	0	1	0	0	51
	4	14	0	0	2	0	0	108	124
	5	11	0	1	0	1	104	4	121
	6	4	116	0	0	0	0	36	156
	7	0	4	0	0	55	0	0	59
	$N_{\mathbf{Y}^\dagger}$	59	144	50	20	57	113	148	

Tabuľka 4.12: Mapovanie zhlukov datasetu $\mathbf{X}^\dagger$ na zhluky $\mathbf{Y}^\dagger$. Riadok tabuľky uvádza číslo zhluku výsledku k-means, $k = 7$ na $\mathbf{X}^\dagger$ a postupne udáva koľko jeho vlastných jedincov sa vyskytovalo v zhlukoch 1 až 7 k-means, $k = 7$ datasetu $\mathbf{Y}^\dagger$. Posledný stĺpec udáva počet jedincov v danom zhluku pre $\mathbf{X}^\dagger$ a posledný riadok naopak počet jedincov v danom zhluku pre $\mathbf{Y}^\dagger$.

Záver

Táto práca pozostávala z troch hlavných častí. V prvej časti sme sa zoznámili so zhlukovou analýzou, jej princípmi a použitím ako klasifikačného nástroja. Uviedli sme jej členenie a charakterizovali vybrané metódy popisom algoritmov a použitia, pričom sme sa sústredili najmä na charakterizáciu metódu k-means a na ňu nadväzujúcich metód. Predstavili sme spôsoby ohodnocovania podobnosti jednotlivých objektov i zhlukov. Zostavili sme teoretický odrazový bod do ďalšieho používania zhlukovania.

Ďalšia časť sa zaoberala implementáciou a použitím zvolených algoritmov k-means a jeho variant na simulovaných dátach. Demonštrovali sme dva spôsoby určenia možného počtu zhlukov a nadviazali ukážkami, kedy túto hodnotu dodržíme, podhodnotíme či nadhodnotíme. Túto časť sme doplnili o experimenty s rôznou voľbou mier hodnotiacich zhlukovanie, ktoré sme porovnávali s cieľom zistiť, či možno nahradiť pôvodný model ohodnocovania, na ktorého princípe si metóda k-means zakladá. Dospeli sme k záveru, že bez zmeny princípu algoritmu nemožno energie bez následkov meniť úplne ľubovoľne, pretože algoritmy minimalizujú vnútrozhlukovú sumu štvorcov vzdialeností objektov k centru.

V poslednej časti tejto práce sme sa zoznámili s dátami poskytnutými Českým egyptologickým ústavom Univerzity Karlovej. Použili sme metódu k-modes a k-means na viacrozmerné diskrétne dáta ľudí podľa titulov získaných z hrobiek ľudí datovaných do piatej a šiestej perióde Starého Egypta a metódu k-means na viacrozmerné diskrétne dáta hrobiek podľa titulov a typov staviva. Pri aplikácii metód na určenie počtu zhlukov sme narazili na problém nejednoznačnosti výsledku. Metóda kolena v porovnaní s dvojrozmernou spojitou úlohou na našom 69-rozmernom respektíve 48-rozmernom diskrétnom priestore nevykazovala žiadne plató a na jej základe nebolo možné určiť počet zhlukov. V ďalšej práci by bolo možné zamerať sa na inú voľbu predpisu zhlukovacích mier s vhodným prispôsobením algoritmov. Počet zhlukov sme určili na základe obrysového grafu a venovali sa porovnaním vzniknutých mnohopočetných zhlukov a zhlukov vzniknutých v priestore so zvolenou vynechanou podmnožinou príznakov. Podarilo sa nám poukázať na výskyt niektorých výrazných zhlukov v piatej aj šiestej perióde, pre ktoré je príznačné, že zdieľajú rovnaké centrá, teda najtypickejších reprezentantov.

Počas písania práce a aplikácie metód na konkrétnu egyptologickú úlohu sme sa stretli s problémom diskrétneho binárneho priestoru, čo nám otvára cestu na túto prácu naviazať skúmaním metód umožňujúcim zospojitenie priestoru respektíve skúmaním rôznych váh príznakov za účelom lepšej separability zhlukov či náhľadom do problematiky redukcie dimenzionality. V rámci ďalšej práce môžeme zvážiť úpravu metódy módov na dvojfázový princíp s detekciou odľahlých prvkov a s tým spojené hlbšie skúmanie použitia teórie grafov v klasifikácii či rozhodovacích stromov.

Literatúra

[1] Hastie, T., Tibshirani, R., Friedman J.H. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer, c2009. Springer series in statistics. ISBN 978-0-387-84857-0.

[2] Kaufman,L., Rousseeuw,P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344. John Wiley & Sons, USA, 2009. ISBN 0-471-73578-7.

[3] Xu, R., Wundch,D.C. *Clustering*. Hoboken, N.J.: Wiley, 2009. ISBN 978-0-470-27680-8

[4] Řezánková, H., Húsek, D., Snášel, V. *Shluková analýza dat*. 2., rozš. vyd. Praha: Professional Publishing, 2009. ISBN 978-80-86946-81-8

[5] MacQueen, J. et al., 1967. Some methods for classification and analysis of multivariete observations. *Proceedings of the Fifth Berkley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA*. vol.1. pp 281-297

[6] Jung, Y.G., Kang, M.S., Heo, J. Clustering performance comparison using K-means and expectation maximization algorithms, *Biotechnology & Biotechnological Equipment*, 28:sup1, S44-S48, 2014, DOI: 10.1080/13102818.2014.949045

[7] Kodinariya, T.M, Makwana, P.R. Review on determining number of Cluster in K-Means Clustering, *International Journal of Advance Research in Computer Science and Management Studies*,vol. 1, issue 6, pp 90-95, 2013

[8] Yuan, C.; Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. J 2019, 2, 226-235.

[9] Arthur, D., Vassilvitdkii, S. k-means++:The Advantages of careful seeding, dostupné z `https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf` [z 26.11.2020]

[10] Wierzchoń, S.T., Kłopotek, M.A. *Modern Algorithms of Cluster Analysis*.volume 34 of Studies in Big Data. Springer Verlag, 2018. ISBN 978-3-319-69307-1

[11] Jiang,M.F., Tseng,S.S., Su,C.M. Two-phase clustering process for outliers detection, *Pattern Recognition Letters*, Volume 22, Issues 6–7, 2001,Pages 691-700, ISSN 0167-8655

[12] Krbálek, M. *Matematická analýza III*, 3. vydání, V Praze: České vysoké učení technické ISBN 978-80-01-04863-4

[13] Anděl, J. *Statistické metody*, Matfyzpress, Praha, 2007, ISBN 80-7378-003-8

[14] Shalizi, C. *Data Mining*. lecture 08 dostupné z `https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf` [z 27.3.2021]

Príloha 1

Táto príloha obsahuje tabuľky spojené s egyptologickou úlohou, ktorá sa nachádza v kapitole 4. Prvou tabuľkou je vymenovanie jednotlivých kategórií titulov s priradeným číslom, pričom v kapitole 4 využívame číselné značenie. V ďalších tabuľkách porovnávame odlišnosti medzi vzniknutými centrami na báze umiestnenia ich centra.

1	Uncategorized	24	King	47	WithTheNameOfGod
2	AdministrativeTitle	25	Patronage	48	ProvisioningOfKing
3	LegalMatters	26	Entertainment	49	Estate
4	Epithet	27	SunTemple	50	WithTheNameOfKing
5	RankTitle	28	EstateHousehold	51	MarginalZones
6	NonRoyalFuneraryCult	29	PrivacyOfKing	52	Settlements
7	PriestlyTitle	30	Land	53	TaxationAndCollection
8	FamilyRelation	31	Provinces	54	Palace
9	ConnectedWithTheGod	32	PersonalApartment	55	RoyalChildren
10	BodyCare	33	PersonalAdornment	56	RoyalWorks
11	GreatHouse	34	PyramidTown	57	MunicipalAdministration
12	PrivyToTheSecret	35	Farming	58	Residence
13	RoyalPyramidComplex	36	Granary	59	ConnectedWithAnIndividual
14	RoyalAffiliation	37	KingsMother	60	Education
15	HonorificTitle	38	Sanctuary	61	Bodyguard
16	Craft	39	ConnectedWithTheFamily	62	TwoHouses
17	Documents	40	ConnectedWithTheKing	63	Priesthood
18	PrivacyOfGreatHouse	41	Highest	64	RegaliaOrnament
19	Redistribution	42	LowerEgypt	65	Storage
20	God	43	UpperEgypt	66	NameOfKing
21	OrganizationOfLabour	44	PrivacyOfPalace	67	ConnectedWithEmployment
22	Treasury	45	ReligiousTitle	68	GreatCourt
23	Physician	46	Provisioning	69	SixGreatCourts

Tabuľka 4.13: Kategórie titulov s priradeným číslom titulu pre úlohu základného datasetu

All: $\mathbf{X}$

1	administrative title	17	municipal administration – residence	33	open court
2	body care	18	physician	34	pyramid complex
3	body care – Great House	19	priestly title	35	rock-cut tomb
4	body care – king	20	privacy of King	36	rock-cut tomb + stone casing
5	craft	21	privy to the secret	37	rubble built mastaba
6	education – residence	22	ranktitle	38	shaft
7	education – royal children	23	religious title	39	stone + mudbrick mastaba
8	entertainment	24	royal affiliation	40	stone family tomb
9	epithet	25	uncertain	41	stone mastaba
10	family relation	26	unlisted cathegory	42	stone mastaba with two niches
11	farming	27	unlisted tomb type	43	stone mastaba with mudbrick coating
12	foreign land	28	transitional type of tombs	44	stone mastaba with open court
13	honorific title	29	half rock-cut tomb	45	stone mastaba with pillared portico
14	honorific title – King	30	mudbrick mastaba	46	temple
15	honorific title – legal matters	31	mudbrick mastaba with rock-cut chapel	47	unknown
16	legal matters	32	nn	48	unobservable

Tabuľka 4.14: Kategórie titulov s priradeným číslom titulu a typy hrobiek s číslom pre úlohu rozšíreného datasetu a maticu všetkých zmienok

Owners: $\mathbf{X}^{\dagger}$

1	administrative title	16	municipal administration – residence	31	pyramid complex
2	body care	17	physician	32	rock-cut tomb
3	body care – Great House	18	priestly title	33	rock-cut tomb + stone casing
4	body care – king	19	privacy of King	34	rubble built mastaba
5	craft	20	privy to the secret	35	shaft
6	education – residence	21	ranktitle	36	stone + mudbrick mastaba
7	education – royal children	22	religious title	37	stone family tomb
8	entertainment	23	royal affiliation	38	stone mastaba
9	epithet	24	unlisted cathegory	39	stone mastaba with two niches
10	family relation	25	unlisted tomb type	40	stone mastaba with mudbrick coating
11	foreign land	26	transitional type of tombs	41	stone mastaba with open court
12	honorific title	27	half rock-cut tomb	42	stone mastaba with pillared portico
13	honorific title – King	28	mudbrick mastaba	43	unknown
14	honorific title – legal matters	29	mudbrick mastaba with rock-cut chapel	44	unobservable
15	legal matters	30	open court		

Tabuľka 4.15: Kategórie titulov s priradeným číslom titulu a typy hrobiek s číslom pre úlohu rozšíreného datasetu a maticu podľa pochovaných

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		4,8	1,2	8,3	17,2	4,8	17,7	1,6	6,0	1,1	1,2	3,6	11,7	1,0	3,1	0,2	4,1	6,2	1,5	1,2	3,4
		58	14	100	206	58	212	19	72	13	15	43	141	12	37	3	49	75	18	15	41
1	21,3	499	4	15	3	0	3	3	3	24	10	3	3	7	6	12	4	4	5	12	4
2	1,0	23	7	10	10	11	10	12	10	15	5	10	12	12	9	9	11	11	8	11	11
3	2,7	63	7	16	2	5	4	4	4	23	11	6	2	8	7	11	1	3	10	11	3
4	0,8	19	11	12	14	15	14	16	14	9	11	12	16	14	13	11	13	15	12	9	15
5	5,7	133	5	14	0	3	2	2	2	23	9	4	2	8	5	11	1	3	8	11	3
6	4,0	93	2	11	5	4	3	5	3	20	6	3	5	3	2	8	6	4	3	8	6
7	2,2	51	4	15	3	4	3	3	3	24	10	5	1	7	4	12	4	2	7	12	0
8	1,6	38	6	17	3	4	3	3	3	26	12	5	3	9	6	14	4	4	9	14	4
9	3,1	72	3	14	2	3	0	2	2	23	9	4	2	6	3	11	3	3	6	11	3
10	17,8	418	3	14	2	3	2	0	2	23	9	4	2	8	3	11	3	3	6	11	1
11	0,9	20	5	14	2	5	4	4	2	21	9	4	4	8	5	11	3	3	8	11	5
12	1,7	40	6	15	1	4	3	3	3	22	10	5	3	9	6	12	2	4	9	12	4
13	11,0	258	5	16	2	3	2	2	2	25	11	4	0	6	5	13	3	1	8	13	1
14	8,0	187	3	14	2	3	2	2	0	23	9	2	2	6	3	11	3	1	6	11	3
15	1,3	31	3	14	4	3	4	2	4	21	9	2	4	6	5	11	5	5	6	9	3
16	0,2	5	3	12	6	7	6	6	6	21	7	6	6	6	5	9	7	5	6	9	5
17	0,5	12	12	17	15	14	13	15	13	16	8	11	15	11	12	14	16	14	11	12	16
18	0,4	9	4	15	3	4	3	1	3	24	10	5	3	9	4	12	4	4	7	12	2
19	4,4	102	4	15	3	4	3	3	1	24	10	3	1	5	4	12	4	0	7	12	2
20	2,6	61	6	13	9	10	9	11	9	14	4	7	11	7	6	6	8	10	5	4	10
21	0,7	17	7	14	2	5	4	4	4	23	11	6	2	8	7	13	3	3	10	11	3
22	0,2	4	8	13	9	10	7	9	7	20	8	9	9	9	6	14	10	8	7	14	10
23	1,4	33	5	16	2	3	2	0	2	25	11	4	2	8	5	13	3	3	8	13	3
24	0,6	14	8	13	9	10	9	11	9	14	6	9	11	9	8	8	8	10	9	2	12
25	1,6	37	6	13	1	4	3	3	3	22	10	5	3	9	6	12	2	4	9	10	4
26	2,7	63	6	15	1	4	3	3	3	22	10	5	3	9	6	10	0	4	9	10	4
27	0,3	8	6	13	3	2	5	5	5	22	10	5	5	9	8	12	4	6	7	12	6
28	1,4	33	3	10	4	7	4	6	4	19	5	6	6	6	1	7	5	5	4	7	5

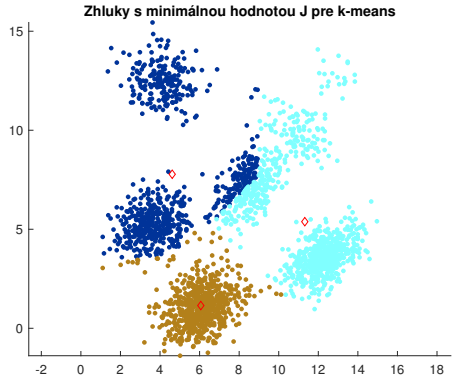
Obr. 4.20: Tabuľka porovnávajúca vzdialenosti centier zhlukov pre výsledky algoritmov k-modes pre $k = 28$ vrámci 5. periódy (stĺpec) a pre $k = 20$ vrámci 6.periódy (riadok). Prvý riadok resp. stĺpec udáva číslo zhluku, druhý udáva percentuálnu veľkosť zhluku a tretí počet objektov v zhluku. Zvyšok tabuľky sú vzájomné vzdialenosti medzi jednotlivými zhlukmi.

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
			9,2	2,1	0,9	2,1	1,8	11,5	5,1	7,8	16,6	14,7	2,7	0,6	0,6	2,9	2,8	0,2	0,0	0,5	4,7	0,8	0,6	3,9	1,4	2,7	0,5	1,0	0,8	1,4
			310	70	30	72	60	387	170	263	557	494	91	20	20	96	94	7	1	16	158	27	20	131	48	90	17	32	26	46
1	0,2	7	14,7	16,4	10,0	12,7	15,0	16,7	14,3	14,7	15,3	15,7	16,3	15,4	15,3	15,4	15,9	15,0	19,4	14,0	15,7	16,7	12,1	16,0	15,7	15,3	18,0	17,0	15,7	16,7
2	0,9	29	4,1	6,0	19,3	5,7	10,9	4,2	5,1	3,9	3,9	5,1	4,9	15,8	18,5	5,0	7,6	6,1	12,2	5,1	4,9	6,1	7,4	6,2	10,2	5,8	12,3	7,3	5,1	4,2
3	0,6	21	9,0	10,3	14,0	7,0	5,6	10,8	9,3	8,9	9,7	9,7	10,6	9,5	12,0	9,4	8,9	10,5	10,5	9,6	9,8	10,6	7,6	6,8	7,1	9,5	12,5	7,7	9,9	10,9
4	1,1	37	10,6	10,6	12,2	8,7	6,2	12,5	10,1	10,5	11,5	11,6	12,5	7,4	8,6	9,6	9,4	12,4	13,7	11,4	11,5	12,6	9,2	9,6	5,4	11,6	13,0	10,2	11,5	12,5
5	3,1	104	2,5	4,5	18,6	4,5	9,5	2,5	3,5	2,5	2,5	3,5	3,5	14,5	17,4	3,5	5,9	4,5	10,5	3,5	3,5	4,5	6,5	4,5	8,5	4,5	14,1	5,4	3,5	1,6
6	0,2	8	24,0	24,3	11,5	22,0	18,8	25,8	23,0	23,8	24,0	25,0	25,0	16,5	16,8	23,3	23,3	23,8	24,8	22,8	24,8	26,0	20,8	21,3	19,3	24,3	19,5	20,5	24,3	25,8
7	1,2	41	2,8	6,5	19,1	4,7	10,3	4,7	5,7	4,8	4,8	2,0	5,7	15,2	18,1	5,6	6,6	6,6	9,8	5,7	5,7	2,9	6,7	7,8	9,4	3,0	12,3	6,8	5,7	4,8
8	0,2	6	6,3	5,3	16,7	6,7	9,0	6,3	7,3	8,3	8,3	7,3	7,3	12,3	14,7	6,3	6,7	6,7	11,0	7,7	7,3	6,3	8,7	10,7	8,7	8,3	13,0	9,7	7,7	8,3
9	22,2	746	2,8	2,5	18,4	4,1	9,6	1,0	3,6	2,5	2,7	3,8	1,7	14,4	17,3	3,5	4,2	2,6	9,1	3,6	1,5	2,8	5,8	5,1	8,7	4,6	15,1	6,4	3,7	2,9
10	17,5	588	1,3	5,0	17,7	3,1	8,8	3,3	4,1	3,3	3,3	0,7	4,3	13,7	16,6	4,1	5,0	5,0	8,2	4,1	4,2	1,6	5,0	6,2	7,8	1,7	14,0	5,2	4,1	3,3
11	0,4	15	11,6	13,1	13,5	10,1	7,7	12,9	12,3	11,1	11,2	12,6	12,2	10,7	9,7	12,1	11,2	13,3	14,2	12,3	12,1	13,6	10,3	9,3	9,3	11,9	11,3	10,8	12,3	12,9
12	22,0	738	2,7	4,6	16,6	4,2	7,7	2,9	3,6	2,6	0,9	3,7	1,9	12,6	15,5	3,6	6,1	4,6	9,1	3,6	3,6	4,7	4,2	3,2	8,7	2,7	14,8	4,3	3,6	2,8
13	4,0	135	4,4	7,0	17,1	4,3	6,1	5,4	7,2	4,5	6,0	5,4	6,0	10,6	13,7	7,0	2,3	7,1	10,0	7,2	4,5	5,4	5,8	5,5	5,4	6,0	11,7	6,2	7,2	6,4
14	0,1	2	21,5	22,5	17,5	19,5	16,5	23,5	20,5	21,5	21,5	22,5	22,5	12,5	16,5	21,5	20,5	23,5	24,5	22,5	22,5	23,5	18,5	20,5	18,5	21,5	16,5	21,5	22,5	23,5
15	0,4	12	18,6	18,6	8,3	16,6	14,4	20,6	17,9	18,6	18,6	19,6	19,6	10,9	12,6	17,6	17,1	19,3	21,1	18,3	19,6	20,6	15,6	17,1	14,3	18,6	18,8	17,8	19,6	20,6
16	6,7	226	3,6	2,9	16,2	3,5	9,0	3,6	2,2	3,6	3,7	4,6	4,6	13,1	15,4	1,9	7,3	3,5	10,1	2,5	4,6	5,6	5,2	6,3	7,3	5,6	15,9	7,2	2,5	3,7
17	0,7	23	11,3	11,5	14,9	10,0	6,3	13,1	11,7	12,0	12,5	12,3	13,3	8,1	11,1	10,7	9,8	12,6	13,3	11,7	12,8	13,1	10,5	10,2	5,9	12,5	10,6	10,4	12,3	13,3
18	0,5	18	8,6	11,9	15,8	8,1	4,6	10,6	10,9	8,7	9,1	9,5	10,1	9,4	11,8	10,9	7,5	11,9	11,3	10,9	9,7	10,5	8,4	6,2	7,1	8,9	9,2	6,8	10,9	10,4
19	10,3	344	2,9	4,7	16,4	2,6	7,9	3,3	3,9	1,3	3,3	3,9	4,3	12,6	15,3	3,7	4,6	4,9	11,0	3,9	2,3	4,9	4,5	3,8	6,8	4,9	13,1	6,3	3,9	3,2
20	0,1	5	17,0	17,8	15,2	15,0	10,6	19,0	17,6	17,0	17,0	18,0	18,0	8,0	14,6	16,8	16,2	18,6	18,0	17,6	18,0	19,0	14,2	14,0	12,4	17,0	16,4	14,2	18,0	19,0
21	0,3	11	14,2	12,5	7,5	12,2	11,5	14,5	11,9	12,5	13,3	15,2	14,3	13,4	10,7	11,5	14,4	12,9	16,8	11,9	13,5	16,2	11,5	11,2	10,6	14,9	16,5	13,5	13,5	14,5
22	0,2	7	18,9	19,7	11,0	16,9	12,3	20,6	17,9	18,9	18,9	19,9	19,6	13,6	8,3	19,0	17,1	20,0	19,9	19,3	19,6	20,6	15,7	16,1	14,1	18,9	14,7	16,6	19,6	20,9
23	0,3	9	11,1	13,1	13,0	10,7	7,1	13,1	11,4	12,7	11,1	12,1	12,1	11,2	11,3	12,1	10,9	12,4	13,4	11,4	13,7	13,1	10,0	11,0	9,2	11,1	11,7	9,6	12,1	12,4
24	0,4	13	12,2	15,5	17,9	12,3	9,1	14,0	14,5	12,6	13,5	13,2	14,5	12,2	14,5	14,5	11,2	15,4	16,1	14,4	13,6	14,2	13,7	11,2	9,8	13,7	4,5	11,4	14,5	14,0
25	0,5	17	11,9	12,1	12,4	9,9	9,6	13,8	12,2	11,8	12,1	12,9	13,1	10,7	6,8	11,1	11,8	13,6	16,4	12,6	12,8	13,9	8,9	12,2	9,6	12,2	17,6	13,4	12,8	13,8
26	0,4	15	10,8	10,8	14,5	8,9	8,3	12,1	10,7	10,3	10,8	11,8	11,8	9,3	12,8	9,8	10,7	12,1	12,5	11,1	11,3	12,8	9,5	7,9	7,4	11,5	14,9	7,5	10,9	12,0
27	0,2	8	7,1	7,6	13,9	6,4	8,1	7,1	9,4	7,1	8,6	8,1	7,6	11,1	13,4	8,6	6,1	8,1	11,9	9,1	6,1	7,1	7,9	8,6	8,9	8,6	13,1	9,6	9,4	9,1
28	1,5	50	3,1	5,3	19,1	5,1	9,3	3,3	4,5	3,4	3,4	4,1	4,2	14,3	15,3	4,5	5,8	5,3	10,5	4,5	4,2	4,9	6,9	4,9	8,4	5,0	14,3	5,3	4,5	3,5
29	3,5	118	3,1	6,7	16,7	4,5	7,5	5,0	5,8	4,7	3,1	2,2	4,0	12,4	15,3	5,8	6,2	6,7	7,0	5,8	5,7	3,1	4,5	4,9	8,6	1,2	14,7	4,1	5,8	5,1

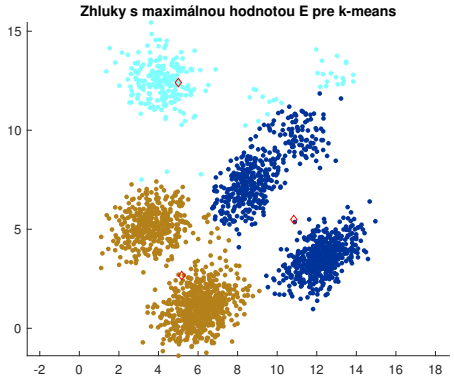
Obr. 4.21: Tabuľka porovnávajúca vzdialenosti centier zhlukov pre výsledky vrámci oboch periód algoritmov k-modes pre $k = 28$ (riadok) a k-means pre $k = 29$ (stĺpec). Prvý riadok resp. stĺpec udáva číslo zhluku, druhý udáva percentuálnu veľkosť zhluku a tretí počet objektov v zhluku. Zvyšok tabuľky sú vzájomné vzdialenosti medzi jednotlivými zhlukmi.

Príloha 2

V tejto prílohe sa nachádzajú sprievodné obrázky riešení zhlukovania na umelých dátach s použitím blokovej a logaritmickej metriky z kapitoly 3.

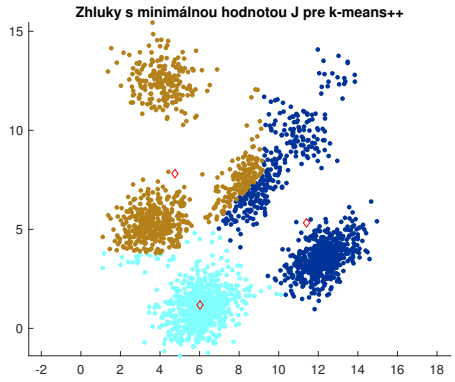


(a)

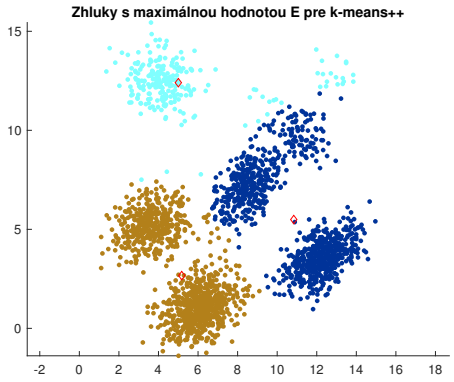


(b)

Obr. 4.22: Výsledky zhlukovacieho procesu k-means, $k = 3$ pre energie uvedené v tab. 3.1 pre blokovú metriku.

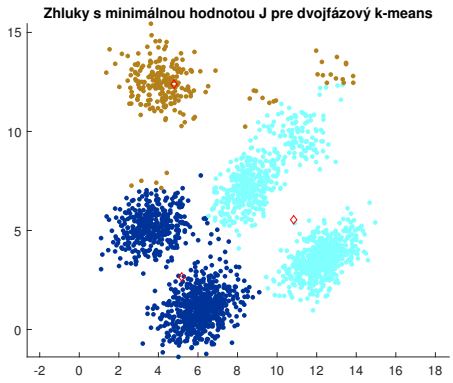


(a)

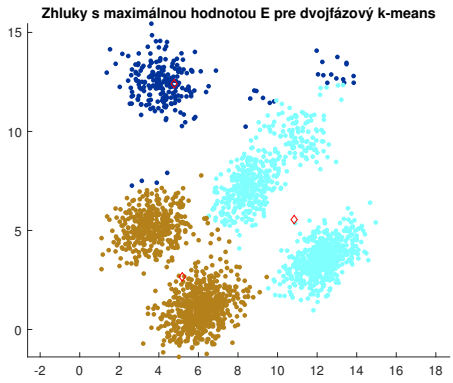


(b)

Obr. 4.23: Výsledky zhlukovacieho procesu k-means++, $k = 3$ pre energie uvedené v tab. 3.1 pre blokovú metriku.

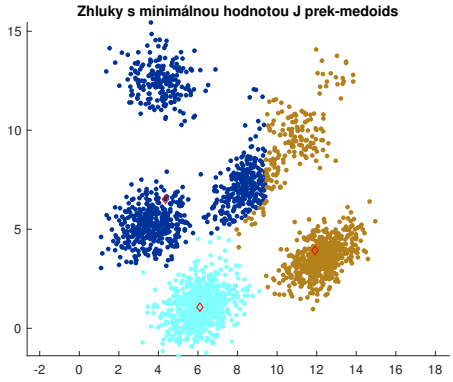


(a)

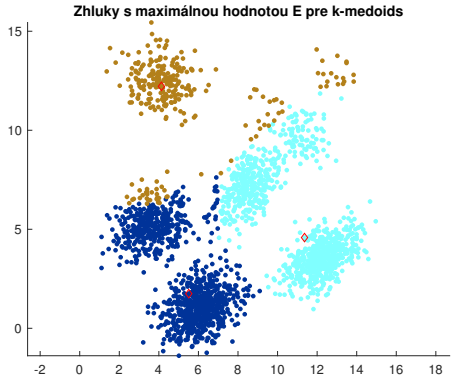


(b)

Obr. 4.24: Výsledky zhlukovacieho procesu dvojfázového k-means, $k = 3$ pre energie uvedené v tab. 3.1 pre blokovú metriku.

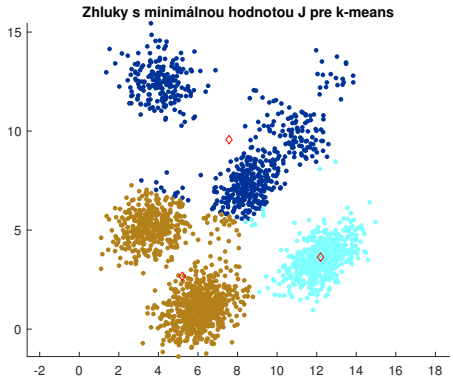


(a)

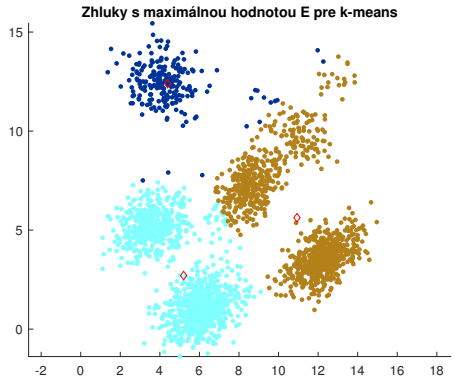


(b)

Obr. 4.25: Výsledky zhlukovacieho procesu k-medoids, $k = 3$ pre energie uvedené v tab. 3.1 pre blokovú metriku.

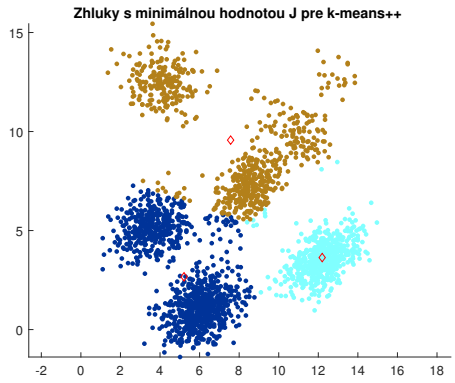


(a)

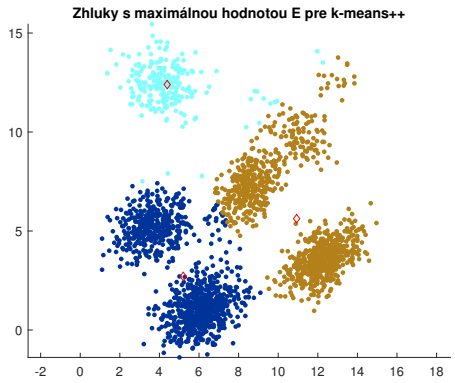


(b)

Obr. 4.26: Výsledky zhlukovacieho procesu k-means, $k = 3$ pre energie uvedené v tab. 3.1 pre logaritmickú metriku.

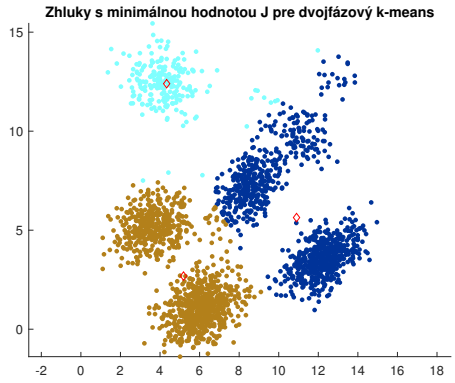


(a)

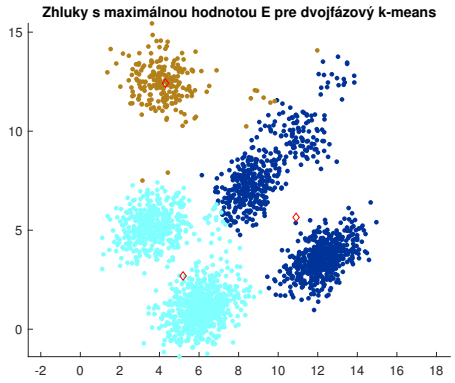


(b)

Obr. 4.27: Výsledky zhlukovacieho procesu k-means++, $k = 3$ pre energie uvedené v tab. 3.1 pre logaritmickú metriku.

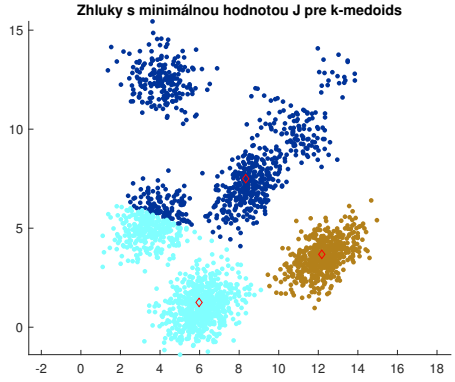


(a)

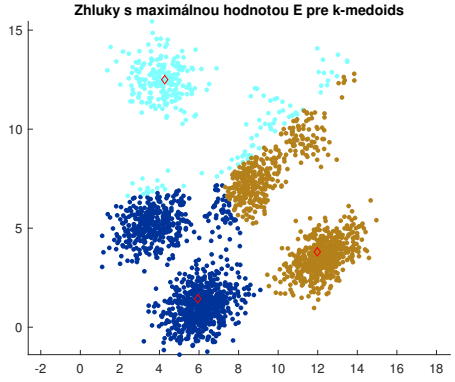


(b)

Obr. 4.28: Výsledky zhlukovacieho procesu dvojfázového k-means, $k = 3$ pre energie uvedené v tab. 3.1 pre logaritmickú metriku.

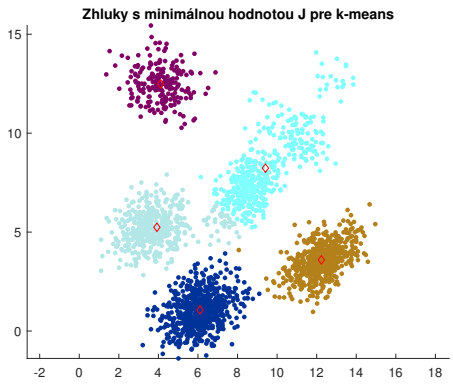


(a)

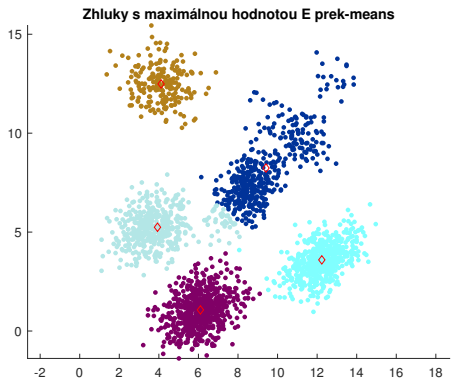


(b)

Obr. 4.29: Výsledky zhlukovacieho procesu k-medoids, $k = 3$ pre energie uvedené v tab. 3.1 pre logaritmickú metriku.

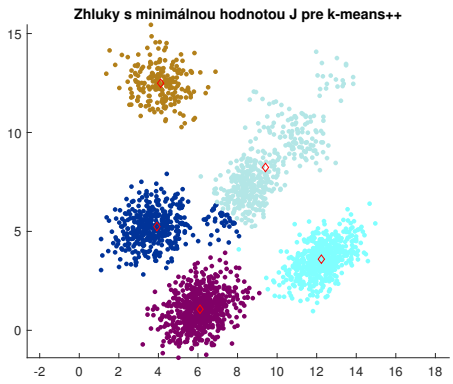


(a)

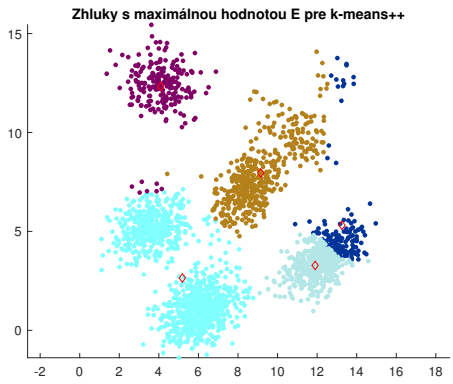


(b)

Obr. 4.30: Výsledky zhlukovacieho procesu k-means, $k = 5$ pre energie uvedené v tab. 3.2 pre blokovú metriku.

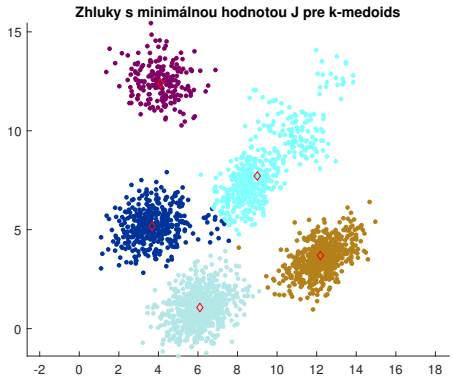


(a)

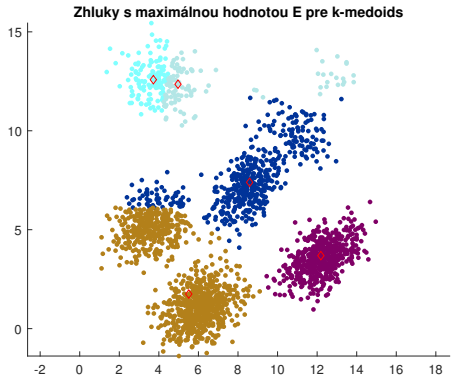


(b)

Obr. 4.31: Výsledky zhlukovacieho procesu k-means++, $k = 5$ pre energie uvedené v tab. 3.2 pre blokovú metriku.

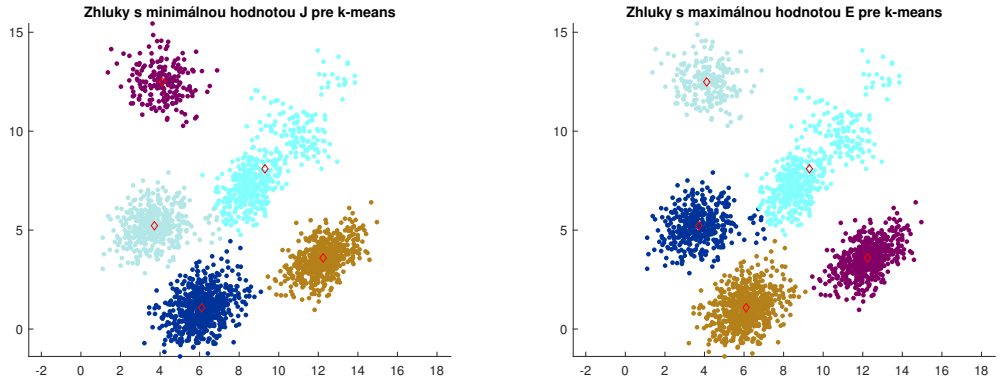


(a)



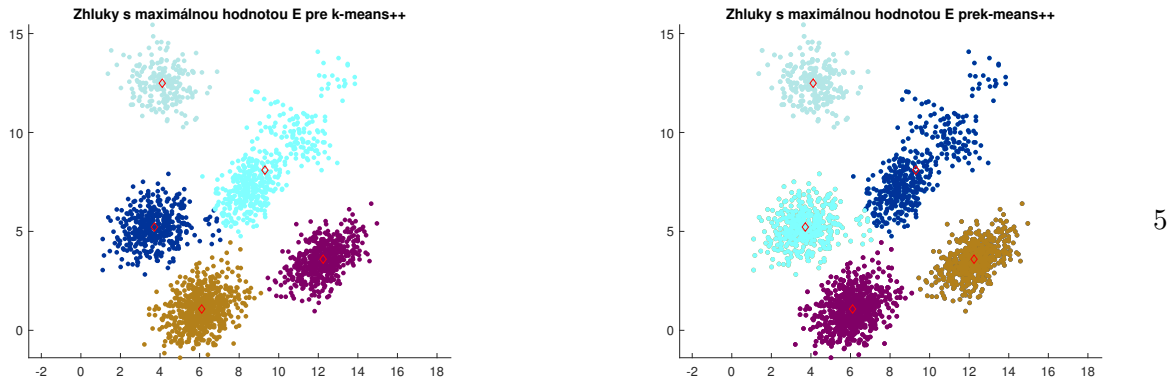
(b)

Obr. 4.32: Výsledky zhlukovacieho procesu k-medoids, $k = 5$ pre energie uvedené v tab. 3.2 pre blokovú metriku.



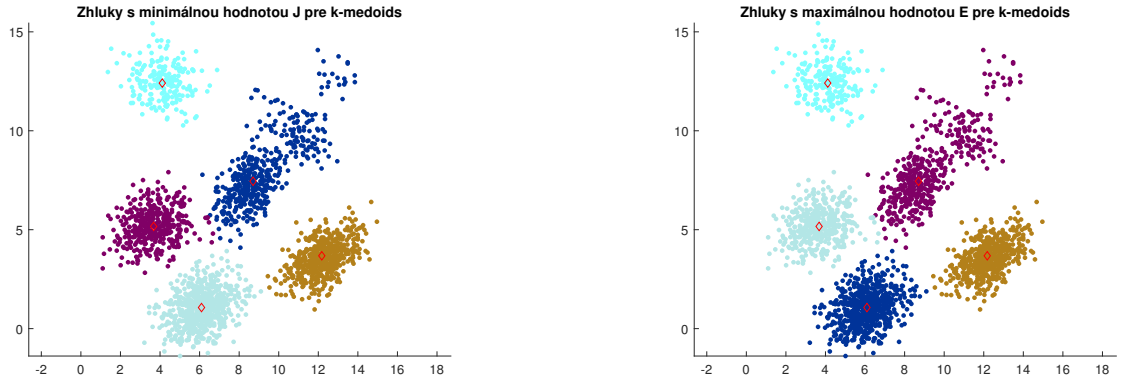
(a) (b)

Obr. 4.33: Výsledky zhlukovacieho procesu k-means, $k = 5$ pre energie uvedené v tab. 3.2 pre logaritmickú metriku.



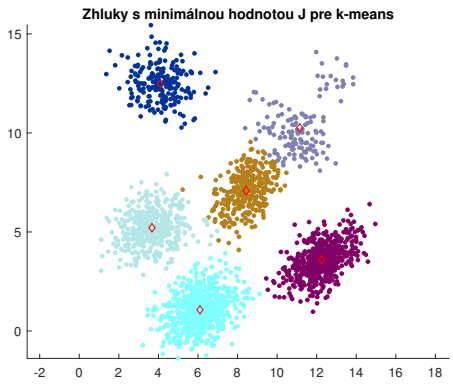
(a) (b)

Obr. 4.34: Výsledky zhlukovacieho procesu k-means++, $k = 5$ pre energie uvedené v tab. 3.2 pre logaritmickú metriku.

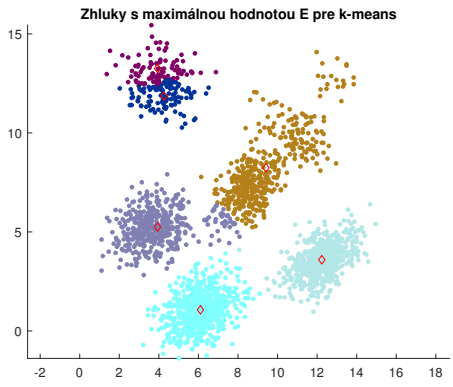


(a) (b)

Obr. 4.35: Výsledky zhlukovacieho procesu k-medoids, $k = 5$ pre energie uvedené v tab. 3.2 pre logaritmickú metriku.

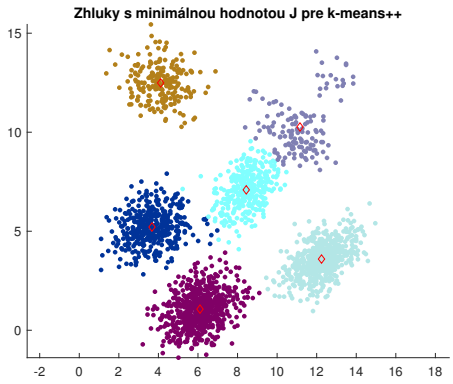


(a)

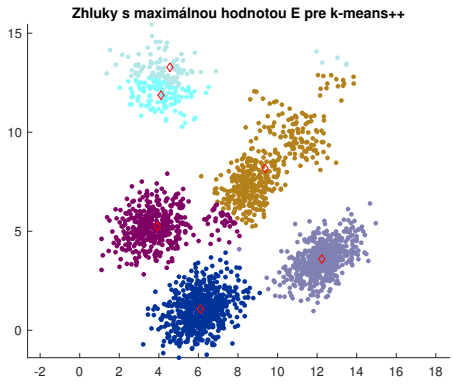


(b)

Obr. 4.36: Výsledky zhlukovacieho procesu k-means, $k = 6$ pre energie uvedené v tab. 3.3 pre blokovú metriku.

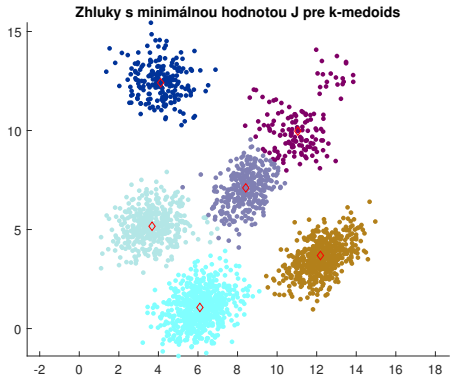


(a)

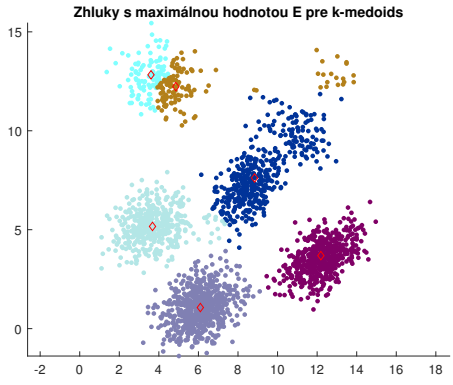


(b)

Obr. 4.37: Výsledky zhlukovacieho procesu k-means++, $k = 6$ pre energie uvedené v tab. 3.3 pre blokovú metriku.

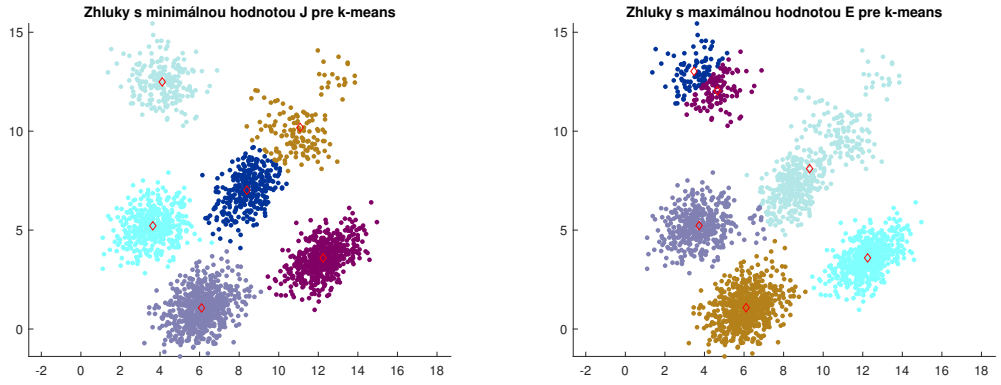


(a)



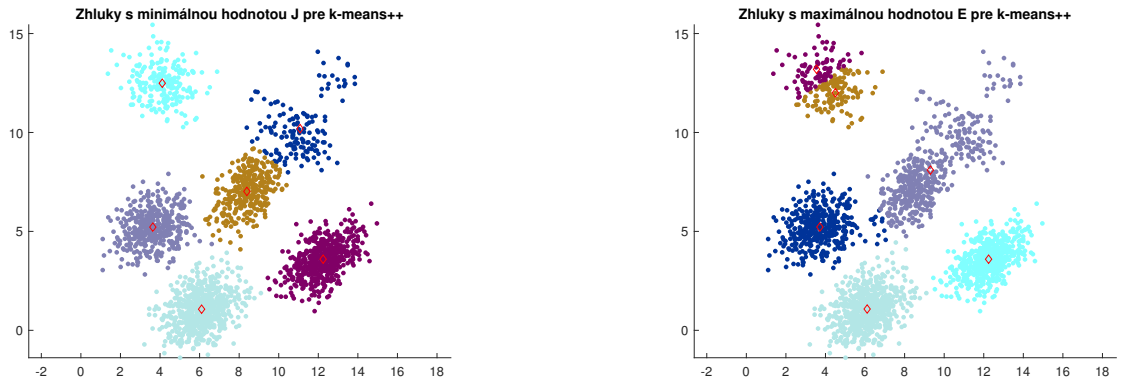
(b)

Obr. 4.38: Výsledky zhlukovacieho procesu k-medoids, $k = 6$ pre energie uvedené v tab. 3.3 pre blokovú metriku.



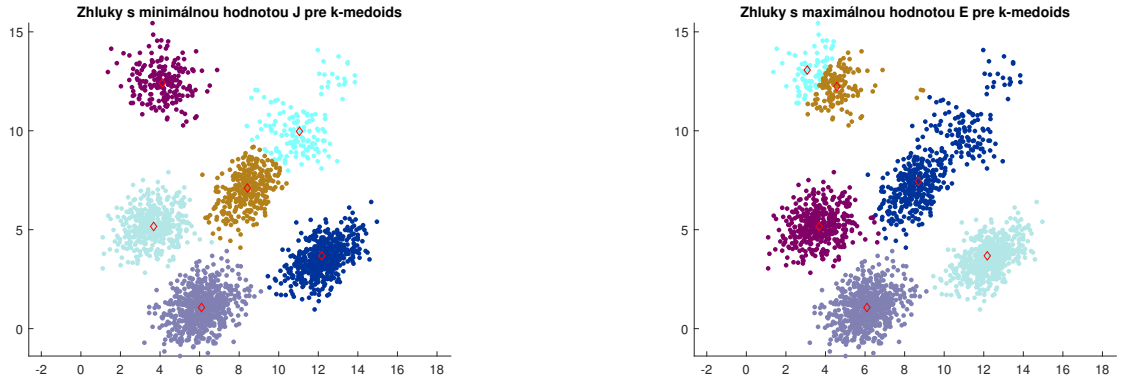
(a) (b)

Obr. 4.39: Výsledky zhlukovacieho procesu k-means, $k = 6$ pre energie uvedené v tab. 3.3 pre logaritmickú metriku.



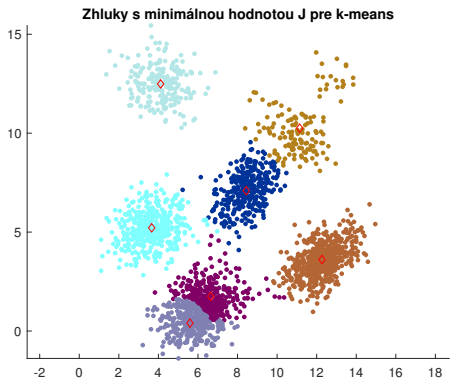
(a) (b)

Obr. 4.40: Výsledky zhlukovacieho procesu k-means++, $k = 6$ pre energie uvedené v tab. 3.3 pre logaritmickú metriku.

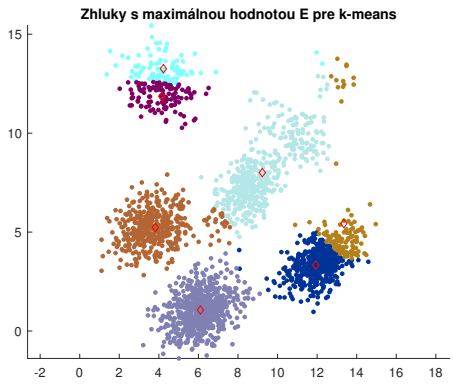


(a) (b)

Obr. 4.41: Výsledky zhlukovacieho procesu k-medoids, $k = 6$ pre energie uvedené v tab. 3.3 pre logaritmickú metriku.

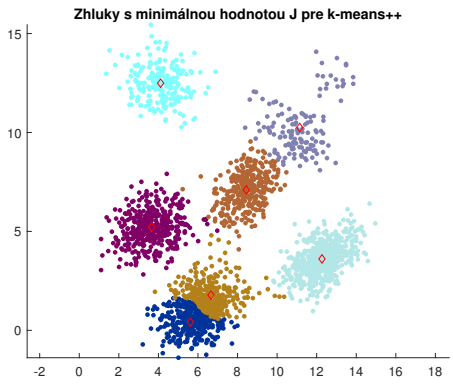


(a)

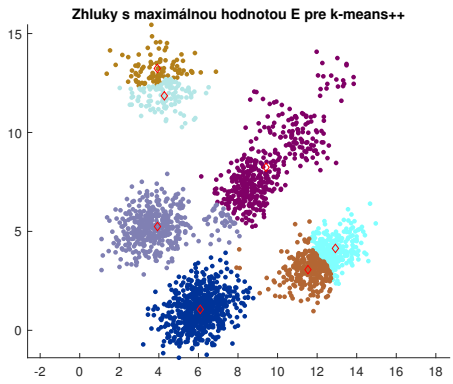


(b)

Obr. 4.42: Výsledky zhlukovacieho procesu k-means, $k = 7$ pre energie uvedené v tab. 3.4 pre blokovú metriku.

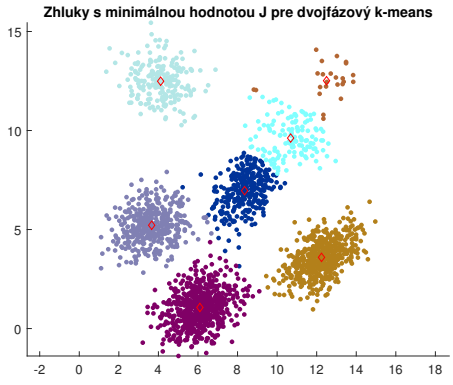


(a)

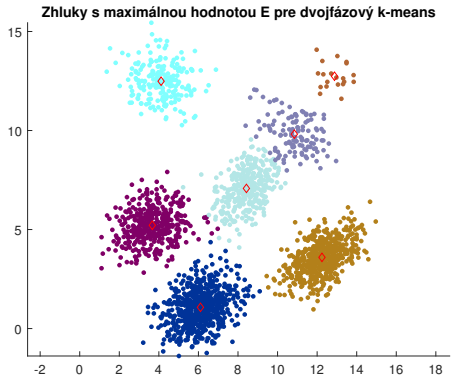


(b)

Obr. 4.43: Výsledky zhlukovacieho procesu k-means++, $k = 7$ pre energie uvedené v tab. 3.4 pre blokovú metriku.

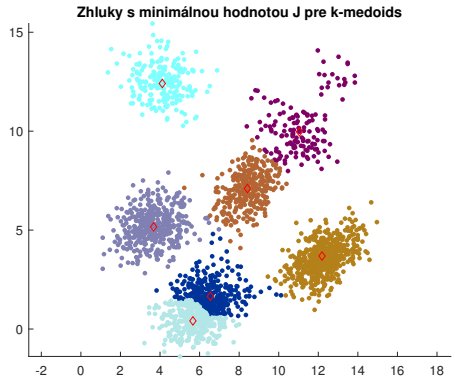


(a)

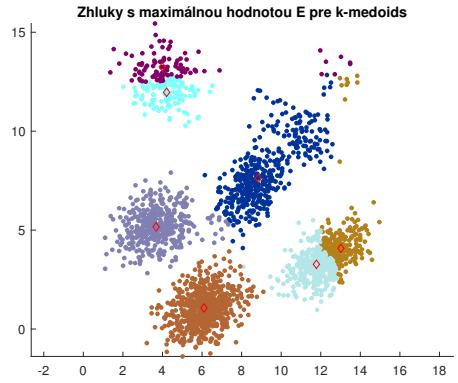


(b)

Obr. 4.44: Výsledky zhlukovacieho procesu dvojfázového k-means, $k = 7$ pre energie uvedené v tab. 3.4 pre blokovú metriku.

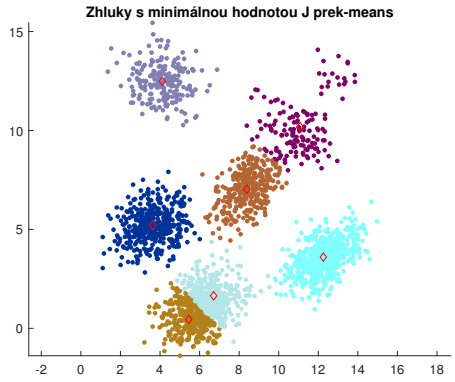


(a)

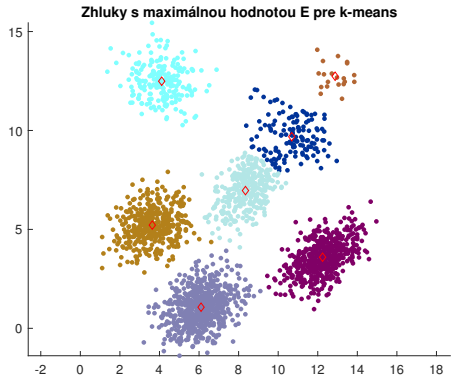


(b)

Obr. 4.45: Výsledky zhlukovacieho procesu k-medoids, $k = 7$ pre energie uvedené v tab. 3.4 pre blokovú metriku.

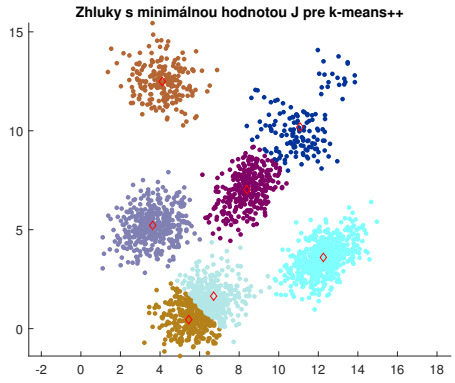


(a)

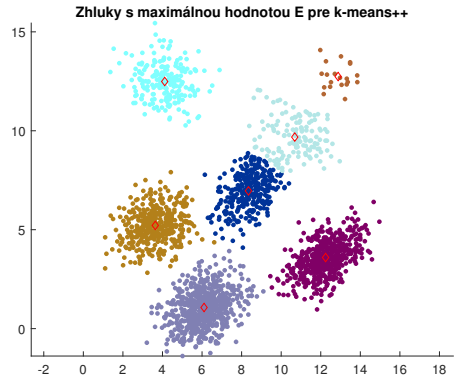


(b)

Obr. 4.46: Výsledky zhlukovacieho procesu k-means, $k = 7$ pre energie uvedené v tab. 3.4 pre logaritmickú metriku.

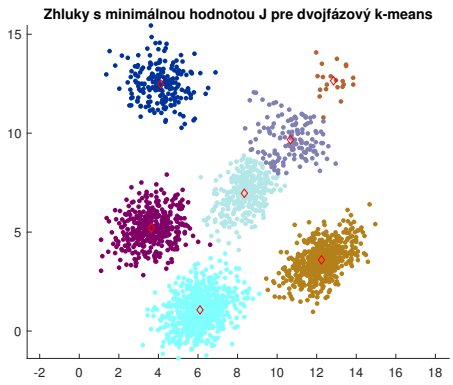


(a)

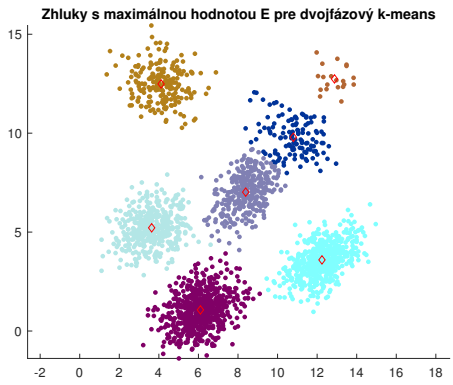


(b)

Obr. 4.47: Výsledky zhlukovacieho procesu k-means++, $k = 7$ pre energie uvedené v tab. 3.4 pre logaritmickú metriku.

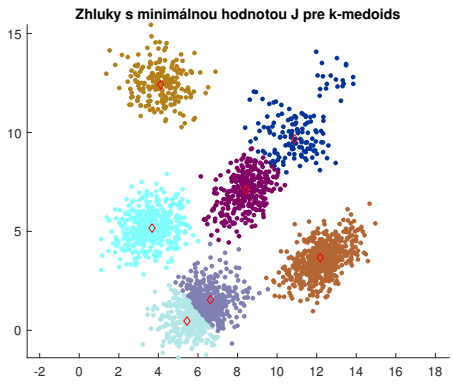


(a)

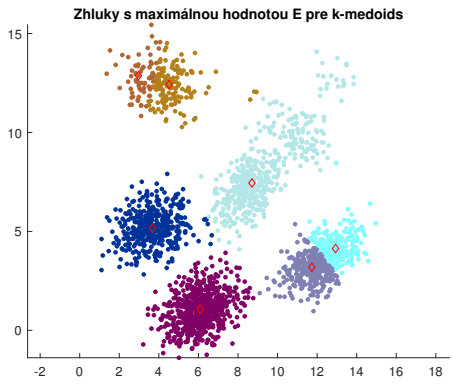


(b)

Obr. 4.48: Výsledky zhlukovacieho procesu dvojfázového k-means, $k = 7$ pre energie uvedené v tab. 3.4 pre logaritmickú metriku.

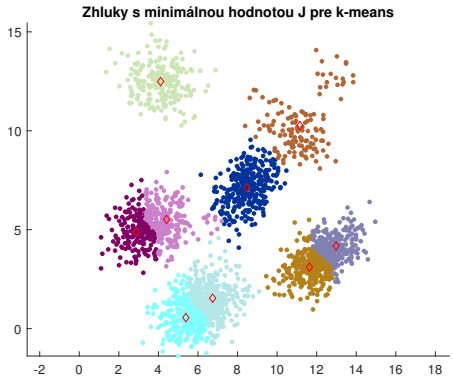


(a)

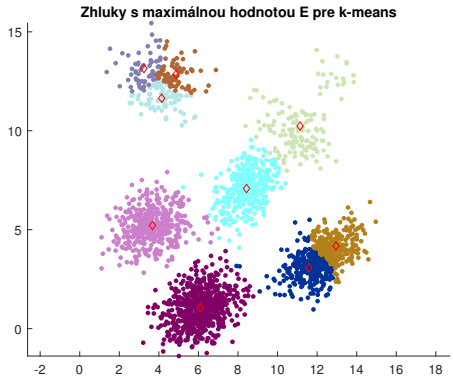


(b)

Obr. 4.49: Výsledky zhlukovacieho procesu k-medoids, $k = 7$ pre energie uvedené v tab. 3.4 pre logaritmickú metriku.

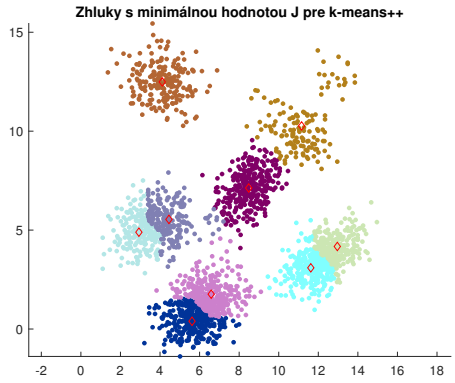


(a)

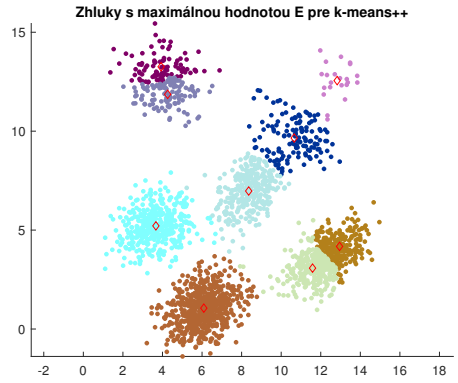


(b)

Obr. 4.50: Výsledky zhlukovacieho procesu k-means, $k = 9$ pre energie uvedené v tab. 3.5 pre blokovú metriku.

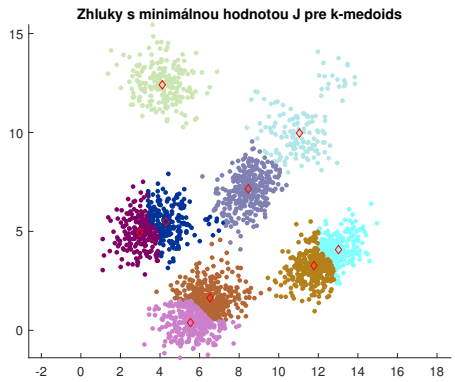


(a)

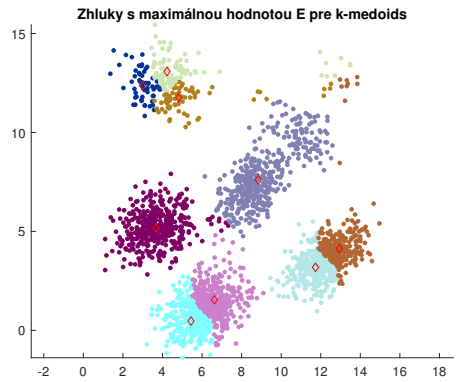


(b)

Obr. 4.51: Výsledky zhlukovacieho procesu k-means++, $k = 9$ pre energie uvedené v tab. 3.5 pre blokovú metriku.

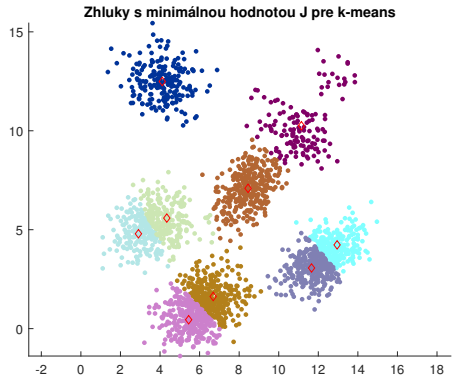


(a)

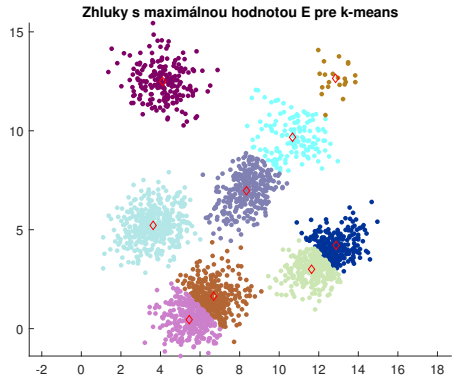


(b)

Obr. 4.52: Výsledky zhlukovacieho procesu k-medoids, $k = 9$ pre energie uvedené v tab. 3.5 pre blokovú metriku.

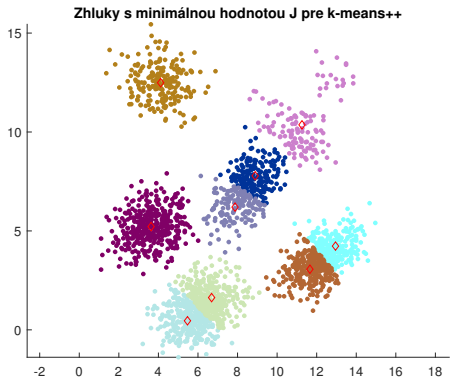


(a)

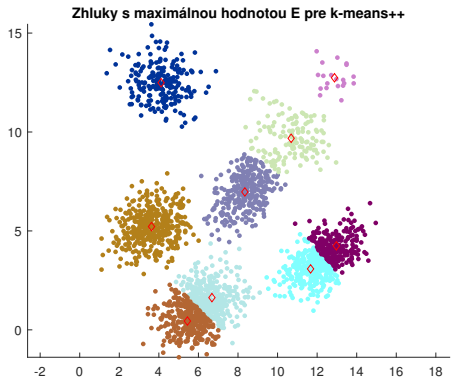


(b)

Obr. 4.53: Výsledky zhlukovacieho procesu k-means, $k = 9$ pre energie uvedené v tab. 3.5 pre logaritmickú metriku.

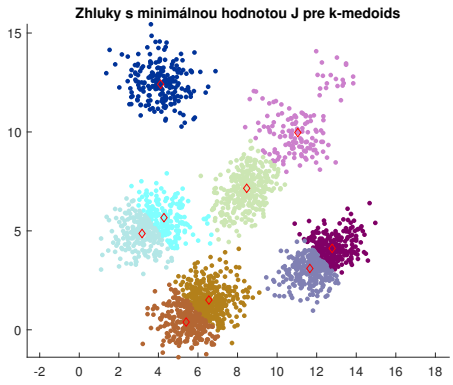


(a)

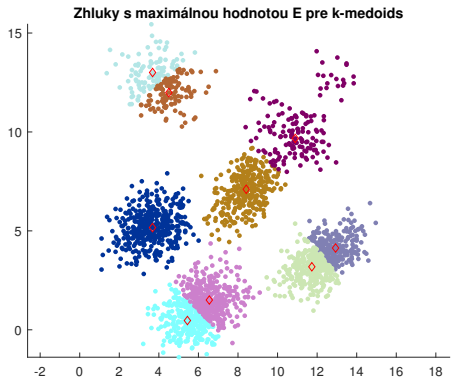


(b)

Obr. 4.54: Výsledky zhlukovacieho procesu k-means++, $k = 9$ pre energie uvedené v tab. 3.5 pre logaritmickú metriku.

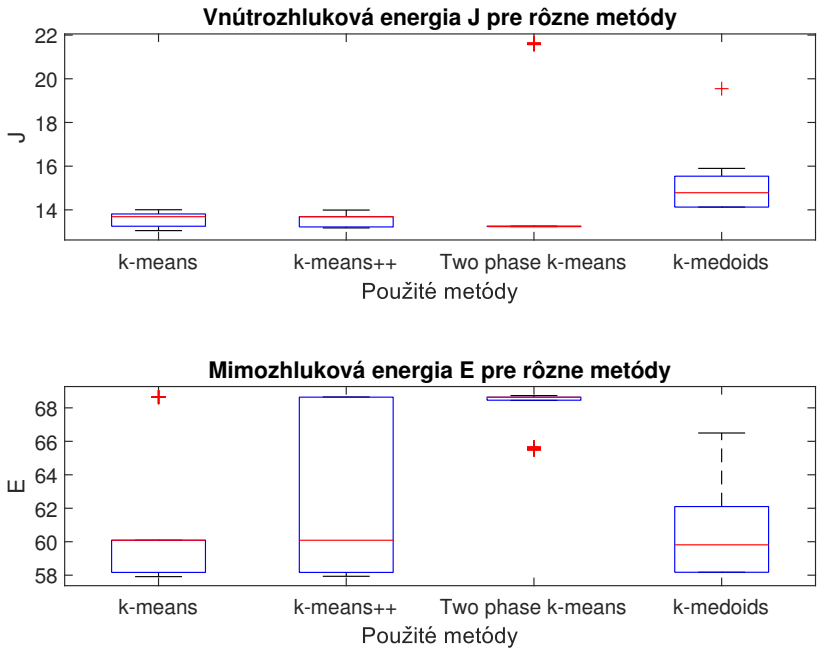


(a)

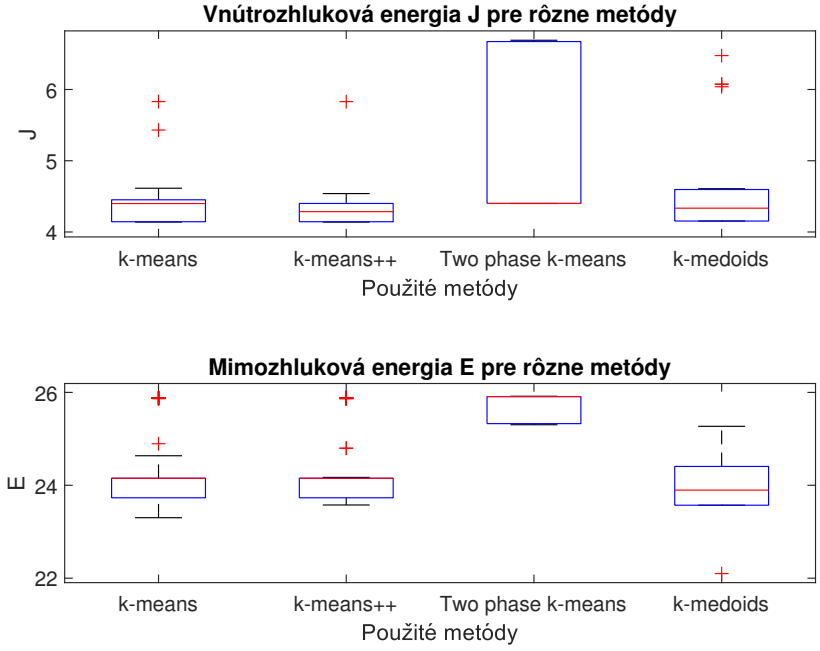


(b)

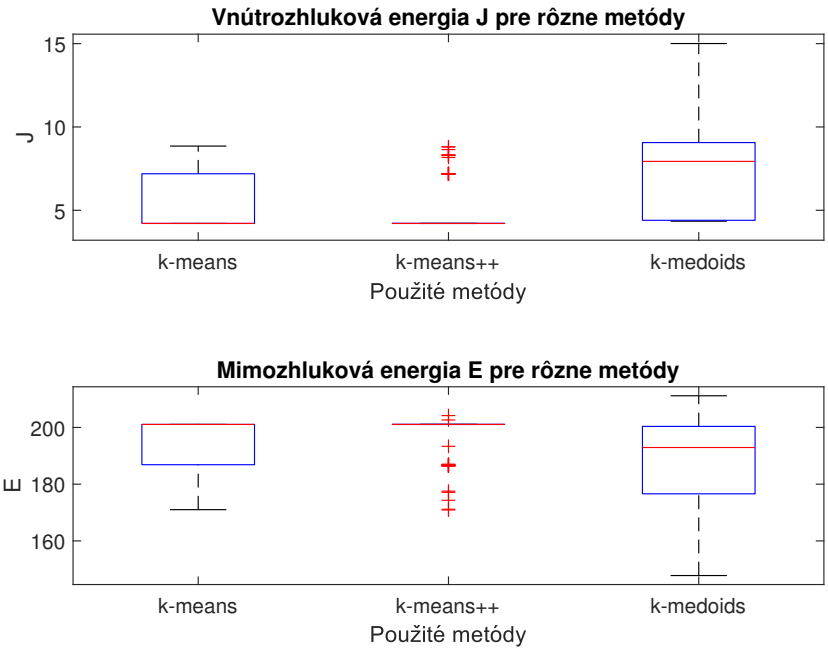
Obr. 4.55: Výsledky zhlukovacieho procesu k-medoids, $k = 9$ pre energie uvedené v tab. 3.5 pre logaritmickú metriku.



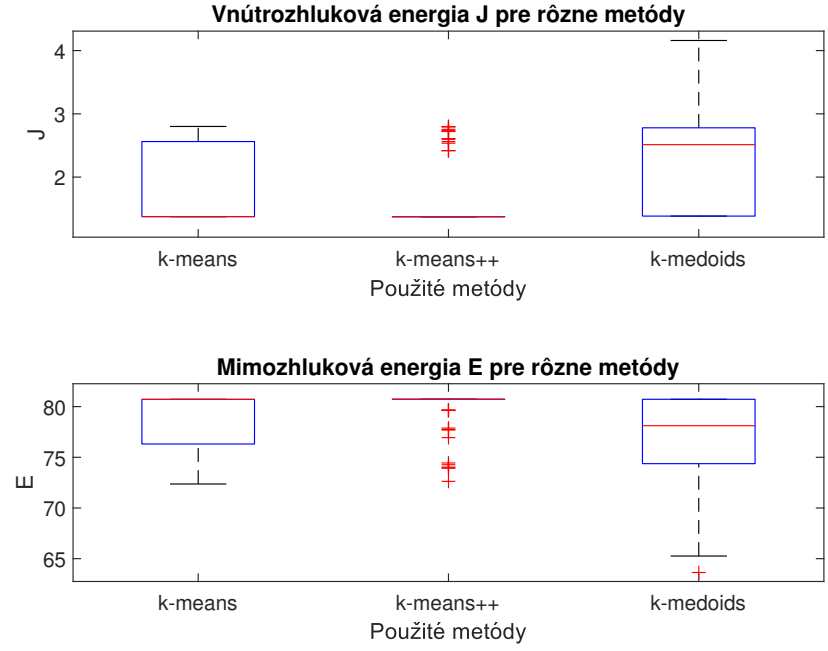
Obr. 4.56: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 3$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s blokovou metrikou.



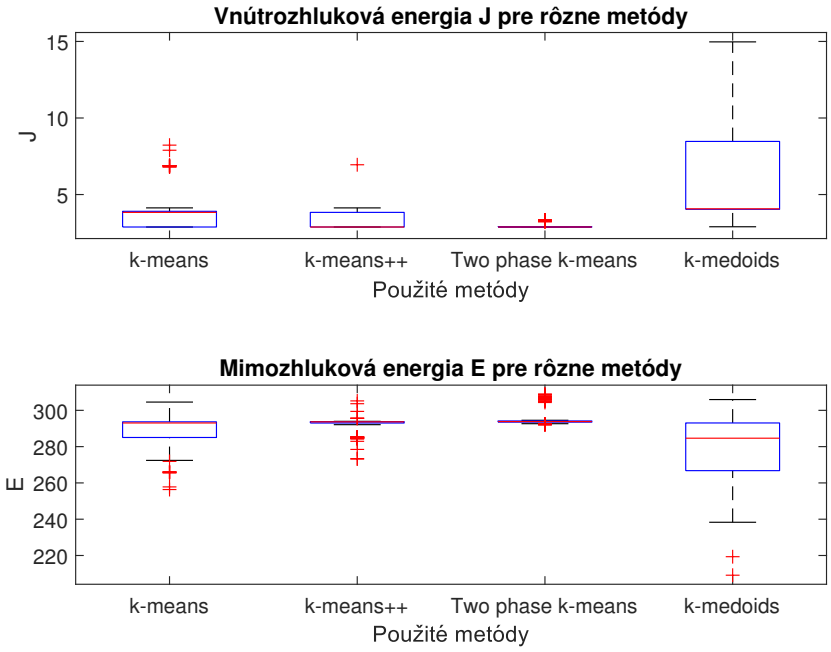
Obr. 4.57: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 3$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s logaritmickou metrikou.



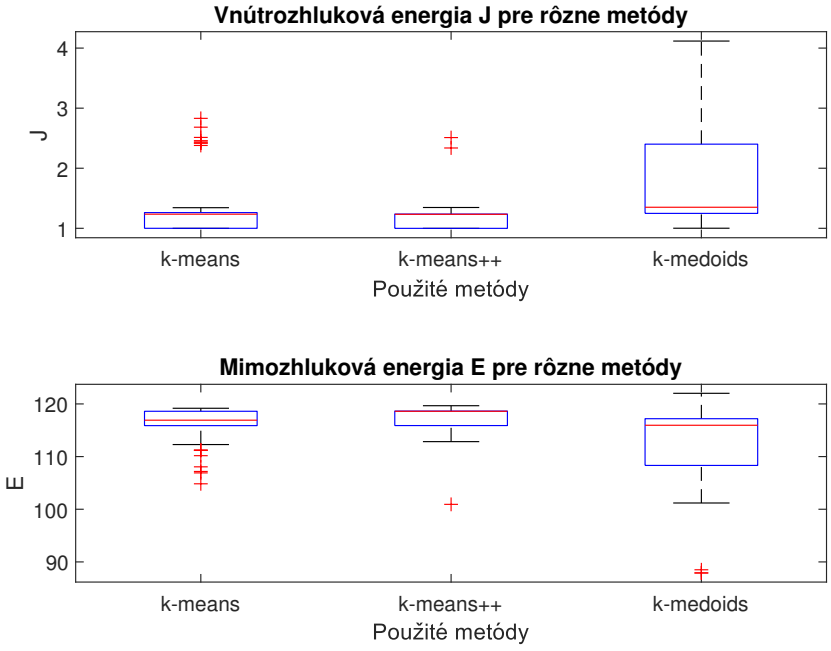
Obr. 4.58: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 5$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s blokovou metrikou.



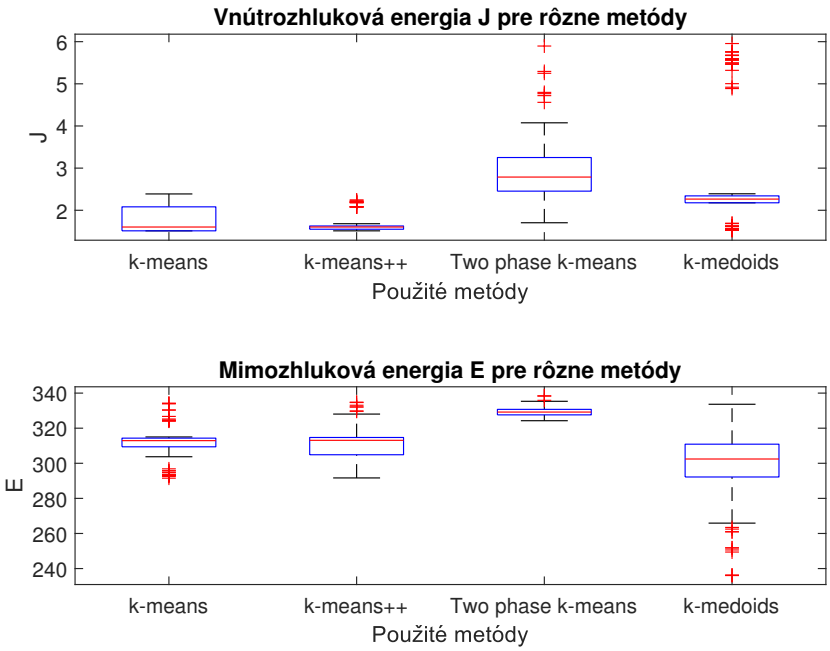
Obr. 4.59: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 5$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s logaritmickou metrikou.



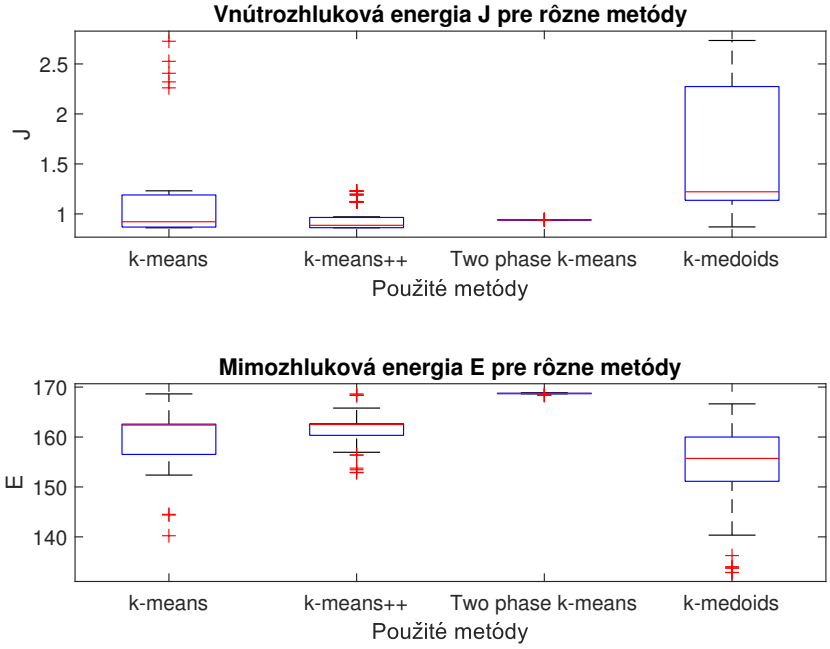
Obr. 4.60: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 6$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s blokovou metrikou.



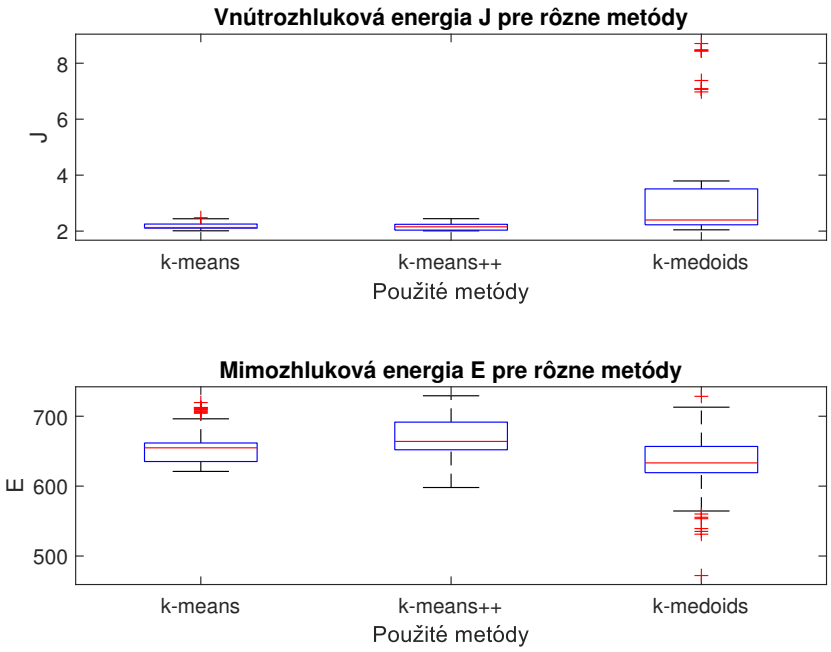
Obr. 4.61: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 6$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s logaritmickou metrikou.



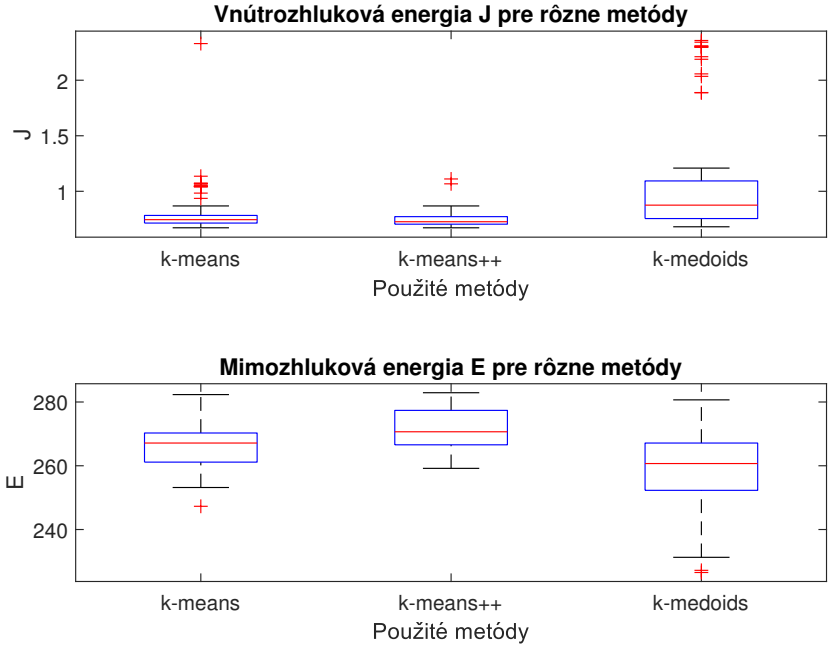
Obr. 4.62: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 7$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s blokovou metrikou.



Obr. 4.63: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 7$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s logaritmickou metrikou.



Obr. 4.64: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 9$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s blokovou metrikou.



Obr. 4.65: Grafické znázornenie vnútrozhlukovej a mimozhlukovej energie metód pre $k = 9$ v priebehu 111 priebehov jednotlivých algoritmov s náhodnými inicializáciami s logaritmickou metrikou.