UNIVERZITA PARDUBICE
FAKULTA EKONOMICKO - SPRÁVNÍ

BAKALÁŘSKÁ PRÁCE

2012 MICHAELA OUŘECKÁ

Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky

Zpracování podkladů pro seminář předmětu PZDM v softwarovém prostředí Clementine - shluková analýza

Michaela Ouřecká

Bakalářská práce
2012

University of Pardubice
Faculty od Economics and Administration

The elaboration of educational materials for the seminar of the subject PZDM in the Clementine software environment - cluster analysis

Michaela Ouřecká

Thesis
2012

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2011/2012**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Michaela Ouřecká**
Osobní číslo:	**E08031**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Informační a bezpečnostní systémy**
Název tématu:	**Zpracování podkladů pro seminář předmětu PZDM v softwarovém prostředí Clementine - shluková analýza**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

1. Data Mining
2. Metodika CRISP-DM
3. Shluková analýza - charakteristika metody
4. Shluková analýza - zpracování konkrétního příkladu v softwarovém prostředí Clementine

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

[1] BERKA, P. Dobývání znalostí z databází. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.

[2] RUD, O.L. Data Mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computer Press, 2001, 329 s. ISBN 80-7226-577-6.

[3] ŘEZÁNKOVÁ, H., HÚSEK, D., SNÁŠEL, V. Shluková analýza dat. Praha: Professional Publishing, 2009, 220 s. ISBN 978-80-86946-81-8

[4] Step-by-step data mining guide [online]. 1.0. c2000 [cit. 2011-06-27]. CRISP-DM 1.0. Dostupné z WWW: http://community.udayton.edu/provost/it/training/documents/SPSS_CRISPWPlr.pdf

Vedoucí bakalářské práce: **Ing. Tomáš Kořínek**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **3. října 2011**

Termín odevzdání bakalářské práce: **30. dubna 2012**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 3. října 2011

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracovala samostatně. Veškeré literární prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu použité literatury.

Byla jsem seznámena s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako Školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Univerzity Pardubice.

V Pardubicích dne 30. 4. 2012 Mgr. Michaela Ouřecká

PODĚKOVÁNÍ:

Tímto bych rád poděkovala svému vedoucímu práce Ing. Tomášovi Kořínkovi za jeho odbornou pomoc, cenné rady a poskytnuté materiály, které mi pomohly při zpracování bakalářské práce.

ANOTACE

Cílem bakalářské práce bylo vypracovat podklady pro předmět PZDM a popsat problematiku data miningu od přípravy dat do konečných výstupů se zaměřením na shlukovou analýzu. Teoretická část zahrnuje seznámení s termínem data mining a popisuje principy vybraných metod shlukové analýzy. Práce uvádí v praktické části příklad zpracování datového souboru prostřednictvím shlukové analýzy v softwarovém prostředí Clementine. Data miningový proces postupuje dle metodologie CRISP - DM.

KLÍČOVÁ SLOVA

Data mining, metodika CRISP-DM, shluková analýza, shlukování dat, analýza shluků, analýza dat

TITLE

The elaboration of educational materials for the seminar of the subject PZDM in the Clementine software environment - cluster analysis

ANNOTATION

The aim of the thesis was to elaborate documents for the seminar of the subject PZDM and describe the problems of data mining in terms of cluster analysis from data preparation to the final outputs. The theoretical part describes the principles of some selected methods of cluster analysis and there is one example from the data file elaborated in Clementine software environment in the practical part of the thesis. Data mining process follow the methodology of CRISP - DM.

KEYWORDS

Data mining, CRISP – DM metodology, cluster analysis, clustering,data analysis

OBSAH

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

ÚVOD

Hlavní náplní, jak již samotný název bakalářské práce napovídá, je přehledně zpracovat podklady pro seminář předmětu PZDM se zaměřením na problematiku shlukové analýzy. Cílem předmětu Základy data miningu je seznámit studenty s metodikami data miningu, metodologií CRISP a základy tvorby modelů. Bakalářskou práci lze rozdělit na dvě části, teoretickou a praktickou.

V teoretické části jsou vymezeny základní pojmy z oblasti data miningu a popsány metody, které data mining využívá. Data mining slouží k nalézání skrytých, prediktivních, dříve neznámých a potenciálně užitečných závislostí v datech. Data mining neboli dolování (dobývání) znalostí z dat je tedy jakousi výstižnou havířskou metaforou. Popularitu si získal i na poli vědy, zejména v matematických oborech. V současné době je v obchodní sféře významnou technologií s velkým potenciálem pro udržení konkurenceschopnosti, protože může poskytnut odpovědi na důležité obchodní otázky. Umožňuje předpovědět chování zákazníka, což podniku dává možnost rozhodovat se efektivně.

Práce si dále klade dále za cíl seznámit čtenáře s metodologií CRISP – DM. Tato metodika byla vyvinuta v rámci Evropského výzkumného projektu. Cílem bylo vytvořit standardní postup pro řešení úloh dobývání znalostí z dat. Na projektu se podílely společnosti NCR (důležitý dodavatel datových skladů), DaimlerChrysler, ISL (autor systému Clementine) a OHRA (významná pojišťovna v Holandsku). Úspěch tohoto projektu tkví v bohatých praktických zkušenostech v oblasti data miningu. [1]

Práce dále uvádí principy vybraných metod shlukové analýzy. Cílem shlukové analýzy je rozdělení souboru dat do relativně homogenních podsouborů. Objekty uvnitř těchto skupin jsou si co nejvíce podobné. Potřeba klasifikovat jevy na základě podobnosti se dostala i do oborů obtížně matematizovatelných jako jsou vědy sociální, lékařské či biologické. Shluková analýza nám umožňuje vytvářet hypotézy o existenci dominantních skupin v množině objektů. [4]

Pro zpracování ukázkového příkladu bude využita nehiearchická shlukovací metoda K means. Silnou stránkou této metody pro seskupování objektů je rychlost a jednoduchost provedení. [9]

Pro vzorový příklad byla vybrána datová sada, jež zahrnuje informace o občanské vybavenosti obcí a počtem obyvatel mezi 5 000 a 10 000 (včetně). Data miniový proces je řešen na základě metodiky CRISP – DM. V praktické části je aplikována metodologie CRISP

– DM na datovém souboru z oblasti občanské vybavenosti obcí celé České republiky s počtem obyvatel mezi 5000 a 10000 (vč.). Data miningový proces proběhl ve specializovaném softwarovém prostředí Clementine firmy SPSS. Společnost SPSS Inc. je významným poskytovatelem softwarových aplikací pro řešení úloh oblasti predikčního modelování. Na trh uvedla první komerční data miningový nástroj, Clementine, v roce 1994. [12] V současné době se jedná o vedoucí produkt mezi analytickými nástroji pro dolování dat. Podniky je proto hojně využíván pro dosažení lepších obchodních výsledků.

Úkolem bylo ukázat práci se souborem a jeho zpracování pomocí shlukové analýzy a na základě získaných výsledků provést zhodnocení rozvoje obcí s počtem obyvatel mezi 5000 a 10000 (vč.) v celé České republice.

1 DATA MINING

Data mining, neboli dolování znalostí z dat, chápeme jako proces hledání skrytých závislostí v datech. [10] V současné době je nezbytnou součástí obchodní praxe v rámci strategie udržení konkurenceschopnosti. Příkladem očekávaného výsledku vytěžování znalostí může být například predikce reagování zákazníka na změny u dodavatelů nebo zjišťování souvislostí mezi nákupním chováním zákazníka pro cílenou nabídku. Data mining se dnes uplatňuje i v oblasti pojišťovnictví, farmaceutickém průmyslu nebo například energetice. [7] Ve sféře informačních technologií se data mining začal rozvíjet jako prostředek pro analýzu obrovských souborů dat, jejichž vznik souvisí s budování datových skladů. Dnes existují snahy navrhnout proces rozhodování, který by napodoboval schopnost zobecňování a tvorby úsudku člověka. V této oblasti jsou využívány též metody umělé inteligence a statistické analýzy. Data mining je pojem mezioborový. Uplatňuje se zde zejména informatika, statistika a oblast, ze které jsou data čerpána. [10]

1.1 Termín data mining

Účelem data miningu je nalezení nové užitečné informace, kterou člověk potřebuje pro rozhodování. Jedná se tedy o nalezení určitých podmnožin dat, které nesou společné vlastnosti. Tyto nalezené podmnožiny by měly být nanejvýš srozumitelné, doposud nepoznané a validní. Nalezený model by měl být platný pro nová data. Termín data mining označuje proces. Začíná od přípravy dat přes hledání podmnožin po vyhodnocení znalostí. Tento proces se často uvádí s přívlastkem netriviální, protože se je založen na posuzování velkého množství výrazů a různých modelů. [10]

1.2 Typické úlohy data miningu

Následující členění podává obraz o základních úlohách dolování dat, které data mining řeší. [1]

- **Klasifikace**

V případě klasifikace se snažíme nalézt znalosti použitelné pro přiřazení nových případů do tříd. [1] Rozumíme jí tedy třídění dat. [3] Tato úloha řeší například klasifikaci klientů banky podle rizikovosti. [1]

- **Predikce**

Při predikci je naším cílem odhadnout ze starších hodnot nějaké veličiny její vývoj v budoucím období. Zásadní roli zde tedy hraje na rozdíl od klasifikace čas. Příkladem použití je předpověď pohybu cen akcií. [1]

- **Sumarizace**

Sumarizace se provádí u velkého počtu dat, abychom měli přehled o jejich struktuře. Využíváme operace sčítání, odečítání a průměrování. Můžeme například počítat s počtem dnů mezi dvěma daty nebo provádět agregaci u ročních součtů v bance. [3]

- **Segmentace**

Segmentace slouží k rozčlenění dat do skupin s podobnými vlastnostmi. [3] Pro segmentaci dat používáme například metodu shlukové analýzy.

- **Deskripce**

Cílem popisu je nalézt skryté vazby v datech nebo jakousi dominantní strukturu v datech. Deskripce údajů by měla být především srozumitelným popisem tříd dat. [1]

- **Hledání „nuggetů“ (vzorů, pravidel)**

Hledání „nuggetů“ znamená nacházení nových, překvapivých znalostí nebo vzorů chování v datech. Příkladem je analýza nákupního košíku (viz oddíl 1.3). [1]

1.3 Aplikační oblasti data miningu

Dobývání znalostí využíváme v řadě oblastí. Mezi nejpoužívanější aplikační oblastni data miningu patří segmentace a klasifikace klientů bank či pojišťoven, jejíž hlavním přínosem je rozpoznání rizikových či solventních klientů. Dále se data mining používá pro predikci vývoje kurzů akcií, spotřeby elektrické energie, analýzu poruch v telekomunikačních sítích, důvodů změny poskytovatele internetu, mobilních operátorů nebo analýzu nákupního košíku. V posledním uvedeném příkladě nás zajímá, zda si zákazník kupuje zároveň některé druhy zboží, například šampón a zubní pastu. Data mining má své uplatnění i při analýze databází pacientů v nemocnici. [1]

1.4 Metody data miningu

Data mining je zastřešujícím pojmem pro mnoho metod (technik modelování). Zde jsou uvedeny některé vybrané analytických procedury:

- **Rozhodovací stromy**

Účelem rozhodovacího stromu je rozdělení dat do odlišných větví tvořících nejsilnější oddělení hodnot závislé proměnné. [7] Data, která jsou zařazena do stejného segmentu, nesou shodné vlastnosti. [5] Můžeme tak například určit segmenty s požadovaným tržním chováním. Nespornou výhodou rozhodovacích stromů jako jedné z metod data miningu je snadná interpretace výstupů. [7] Tato metoda je založena na různých algoritmech. Mezi nejdůležitější patří CHAID (Chi – squared Automatic Interaction Detector), CART (Classification and Regression Trees), QUEST (The Quik, Unbaised, Efficient Statistical Tree) a C5.0. [5]

- **Genetické algoritmy**

Metoda genetických algoritmů je založena na biologickém evolučním procesu „přežití nejadaptabilnějšího“. V praxi to znamená srovnávání většího počtu modelů a jejich úpravu. Následuje nalezení nejvhodnějšího modelu pro danou úlohu. Každému vybranému modelu je přiřazena váha, která vypovídá o jeho schopnost vyhovět našim požadavkům, například předpovídat zůstatky na účtech. Váha určuje šanci modelu na přežití v dalším testování modelů. Modely jsou dále podrobovány operacím jako náhodná změna znamének a hodnot. V termínech evoluční biologie bychom tedy mohli hovořit například o mutaci a klonování. Optimální model získáme po mnoha iteracích (generacích). Tato metoda je stále více oblíbená v důsledku zvyšující se výkonové efektivity dnešních počítačů. [7]

- **Neuronové sítě**

Tato metoda stojí na principu fungování lidského mozku, který je schopen přijímat informace a poučit se ze zkušenosti. Nejprve zde dochází k rozdělení dat na trénovací a testovací množinu. Potom se vstupním neuronům přiřadí váhy. Podstatou je porovnávání vstupů se skutečnou hodnotou, jehož výstupem je chyba, na základě které se původní váhy upraví. Proces „učení“ končí dosažením předem stanovené minimální chyby. Nevýhodou neuronových sítí je nesnadná interpretace výsledků. [7] Kohonenovy mapy patří mezi nesupervidované neuronové sítě. Více se o metodě Kohonenových map zmíníme v oddíle 4.4.1.1. [11]

- **Lineární regrese**

Technika lineární regrese, stejně jako metoda logistické regrese, využívá statistické metody. Účelem této metody je kvantifikace závislosti mezi dvěma spojitými proměnnými, závislou a nezávislou. Závislou proměnnou se snažíme predikovat. Nezávislou proměnnou můžeme

označit jako prediktivní. Podstatou je nalezení takové přímky procházející mezi jednotlivými body, pro kterou platí, že součet druhých mocnin odchylek od každého bodu je minimální. [7]

- **Logistická regrese**

Hlavní rozdíl mezi technikou lineární a logistické regrese spočívá v tom, že logistická regrese pracuje s diskrétní (kategoriální) závislou proměnnou. Nejedná se tedy o spojitou proměnnou, jak tomu bylo v případě lineární regrese. Jinak jsou si obě zmíněné metody velmi podobné. [7]

- **Shluková analýza**

Jedná se o rozdělení dat do skupin s podobnými vlastnostmi. [5] Metodou shlukování se budeme zabývat více v kapitole 3.

1.5 Metodiky

Metodiky poskytují jednotný rámec pro řešení úloh data miningu. Ukázkový příklad v praktické části bude realizován podle metodiky CRISP – DM, která na rozdíl od metodiky 5A nebo SEMMA není „softwarově závislá“. [1]

1.5.1 Metodika 5A

Název je zkratkou pro fáze, které při řešení úlohy dobývaná znalostí procházíme.

- Assess – zhodnocení požadavků projektu,
- Access – shromáždění dat,
- Analyze – analýzy,
- Act – přeměna znalostí na akční znalosti, tzn. obsahující rozhodnutí, výsledek,
- Automate – uplatnění výsledků analýz v praxi. [1]

1.5.2 Metodika SEMMA

Metodika nese jméno, které je akronymem pro jednotlivé etapy procesu dobývání znalostí.

- Sample – výběr objektů patřičně vhodných pro naši úlohu,
- Explore – explorace dat,
- Modify – provádění datových transformací,

- Model – analýza dat pomocí vhodných modelovacích technik,
- Assess – interpretace. [1]

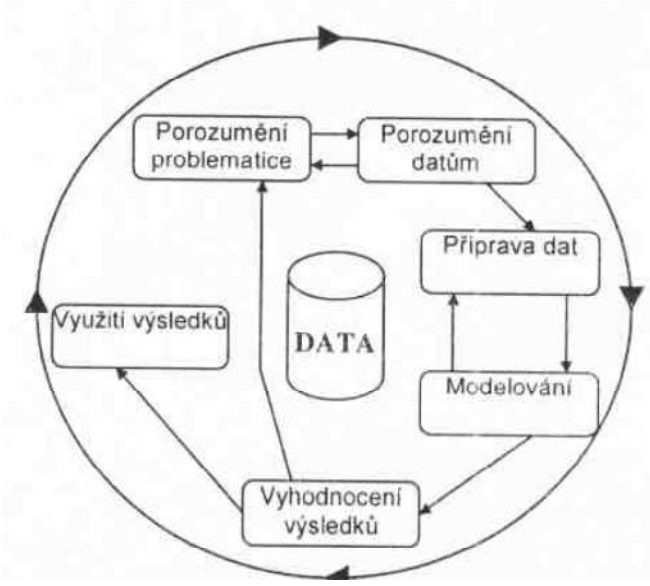
1.5.3 Metodika CRISP – DM

Metodika CRISP –DM (CRoss – Industry Standard Process for Data Mining), životní cyklus projektu dobývání znalostí a náplň jednotlivých kroků je popsán v kapitole 2.

2 METODIKA CRISP – DM

Metodika CRISP – DM byla vyvinuta konsorciem firem v rámci Evropského výzkumného projektu, jehož účelem byl vznik standardního procesu dolování znalostí z databází, který by umožnil řešit data miningové úlohy efektivněji a levněji. [1] Její vznik je datován k roku 1996. V té době byla společnost Deimler-Benz [12], která spolupracovala na Evropském projektu [1], před ostatními firmami výrazně napřed v aplikování data miningu pro své obchodní záměry. Brzy začal být o techniku dolování dat na trhu obrovský zájem. [12] Mezi další společnosti, které se podílely na Evropském projektu, patřila firma NCR (přední dodavatel skladových projektů) nebo OHRA (holandská pojišťovna). [1] Metodika CRISP tedy nevznikala na teoretické akademické půdě. Její úspěch tkví možná právě v praktické zkušenosti s data miningem v reálném světě. [12]

Vzhledem ke skutečnosti, že zpracování příkladu se drží této metodologie, jsou jednotlivé části rozděleny do šesti etap metodiky CRISP – DM. Pořadí kroků není pevně stanoveno. Závisí na výsledku v jedné fázi. Navíc se dle potřeby můžeme k některým etapám vracet. Na Obrázku 1 šipky tvořící elipsu značí cyklus procesu dolování znalostí. [1]



Obrázek 1: Fáze CRISP – DM

Zdroj: [1]

2.1 Porozumění problematice

První etapa je zaměřena na pochopení cílů projektu. Naše požadavky musíme formulovat tak, aby byly řešitelné data miningem. Kromě určování požadavků na řešení zahrnuje tato

etapa posouzení rizik, nákladů, výčet lidských i výpočetních zdrojů a předběžné naplánování práce. [1]

2.2 Porozumění datům

Tato fáze zahrnuje sběr dat, který je následován bližším seznámení s daty. Jedná se zejména o vytipování možných podmnožin záznamů, které nám umožní první vhled do souboru. Posuzujeme i kvalitu nasbíraných dat a identifikujeme případné nedostatky. V rámci statistické analýzy využíváme zejména deskriptivní charakteristiky dat jako četnost, průměr, minimum, maximum atd. [1]

2.3 Příprava dat

Předzpracování dat spočívá v posouzení a výběru relevantních proměnných pro naši analýzu. Součástí přípravy dat je často standardizace dat. Pro získání kvalitních výsledků je potřeba identifikovat odlehlé objekty. [8] Datové soubory nemusí být vždy úplné a může se stát, že některé údaje z nejrůznějších příčin chybí. Hodnotu například nebylo možné změřit nebo je zjištěná hodnota nesmyslná. Při předzpracování dat v rámci shlukové analýzy můžeme objekt s chybějícím údajem ignorovat nebo chybějící údaj nahradit. [7]

2.4 Modelování

Ve fázi modelování pracujeme s daty pomocí zvolených data miningových analytických metod (algoritmů pro dobývání znalostí). Pro dosažení kvalitních výsledků bychom obvykle měli kombinovat více různých algoritmů, například shlukování a rozhodovací stromy. Následuje ověření námi zjištěných znalostí z pohledu metod dobývání znalostí. V případě klasifikačních znalostí, které umožňují například rozpoznat klienty neplatící úroky, se může jednat o jejich testování na nových (nezávislých) datech. [1]

2.5 Vyhodnocení výsledků

Po aplikaci analytických algoritmů provedeme interpretaci výsledků a uvažujeme o způsobu využití získaných znalostí. Získané výsledky je třeba posoudit z manažerského úhlu pohledu s ohledem na účel úlohy, který jsme formulovali v první fázi data miningového procesu. [1]

2.6 Využití výsledků

Závěrečná etapa je věnována „transformaci“ získaných výsledků do podoby stravitelné a využitelné pro zadavatele úlohy (manažera), který provádí aplikaci výsledků. Výstupem může

být například závěrečná správa nebo zavedení nového systému na softwarové nebo organizační úrovni. [1]

3 SHLUKOVÁ ANALÝZA

Shluková analýza je jednou ze základních metod dobývání znalostí. Principem shlukové analýzy je tvorba skupin objektů. Dva objekty zařazené do stejné skupiny jsou si více podobné než dva objekty z různých shluků. Podobnost dvou objektů je stanovena na základě množiny vlastností, kterými je objekt charakterizován. [8]

V literatuře se objevují pojmy učení s učitelem (supervised learning) a učení bez učitele (unsupervised leasing). V prvním případě je předem dána příslušnost objektů do známých skupin. Cílem učení s učitelem je vytvořit model, podle něhož by bylo možné objekty, u kterých není známa informace o příslušnosti ke skupině, zařadit do daných skupin. Shlukování je jednou z metod učení bez učitele. Předem neznáme příslušnost objektů k jednotlivým skupinám. Obvykle neznáme ani počet skupin. Úloha shlukování bez učitele si klade za cíl klasifikovat všechny objekty zahrnuté do analýzy. (Pojem klasifikace zde není užíván pro označení metody vycházející z principu učení s učitelem.) [8]

3.1 Prvky shlukování

Vstupem pro analýzu je datová matice. V případě analýzy výběrového šetření je počet objektů obvykle označován písmenem n. Prvky vektoru pozorování jsou hodnoty proměnných (vlastností, znaků, atributů). V datové matici budou sloupce odpovídat jednotlivým proměnným. Vstupní matice o rozměru n x m je označována písmenem X a její prvky x_{ij}, kde i = 1, 2, .., n a j = 1, 2,..., m. Vektor hodnot znaků budeme nazývat záznam. Objekty (záznamy) budeme zapisovat do řádků. Termínem shlukování je obvykle myšleno spojování objektů do shluků. Při shlukování objektů tedy sledujeme podobnost vektorů, které tvoří řádky matice. V oblasti vyhledávání informací se můžeme též setkat se shlukováním proměnných. Další možností je současné shlukování proměnných i objektů. [8]

3.2 Metriky

Jak již bylo zmíněno výše, snažíme se určit podobné skupiny objektů takovým způsobem, aby byl objekt jedné skupiny co nejvíce podobný objektům stejného shluku a co nejméně objektům v jiných shlucích. Míry podobnosti nabývají hodnot od nuly (maximální odlišnost) do jedničky (shodnost). Segmentační analýza však v běžné praxi používá míry nepodobnosti (vzdálenosti). Dvojice objektů je charakterizována vzdáleností. Čím je vzdálenost mezi těmito objekty menší, tím jsou si podobnější. [8]

Jako míra podobnosti objektů (míru vzdálenosti objektů) můžeme využít například Euklidovskou metriku (vzdálenost), která patří mezi nejznámější. Vychází z geometrického modelu dat. Objekty jsou charakterizovány p znaky. Objektům přiřadíme body p – rozměrného euklidovského prostoru $E_{p.}$ Podobnost objektů je tím větší, čím je vzdálenost jejich bodů (metrika) menší. [4]

Pro Euklidovskou metriku d bodů $A = (a_1, a_2, ...a_p)$, $B = (b_1, b_2, ...b_p)$ platí

$$d(A,B) = [\sum_{i=1}^{p} (a_i - b_i)^2]^{1/2}. \quad (1)$$

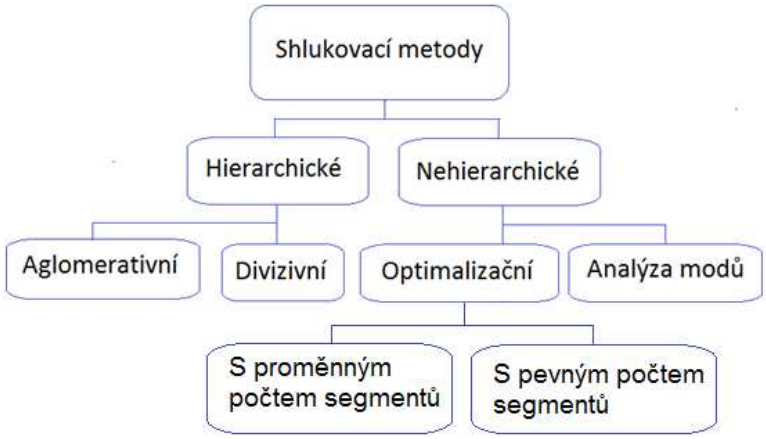
Vzdálenost dvou bodů A, Bj z Ep je obecně definována jako nezáporná reálná funkce d(A,B), pro kterou platí:

1. $d(A,B) = 0 \Leftrightarrow A = B$
2. $d(A,B) \geq 0$ pro všechny body A, B, z E_p
3. $d(A,B) = d(B,A)$
4. $d(A,C) \leq d(A,B) + d(B,C)$ pro každou trojici bodů A, B, C z $E_{p.}$

Čtvrtá podmínka je známa jako trojúhelníkovitá nerovnost. [4]

3.3 Metody shlukové analýzy

Termín shluková analýza je označením pro celý soubor metod. Cílem shlukování může být tvorba hierarchie shluků nebo zařazení objektů do daného počtu shluků. Shlukovací metody se liší dle postupu shlukování na hierarchické a nehiararchické (metody rozkladu). Následující diagram znázorňuje podrobnější dělení shlukovacích metod. [8]



Obrázek 2: Rozdělení metod shlukové analýzy

Zdroj: upraveno na základě [9]

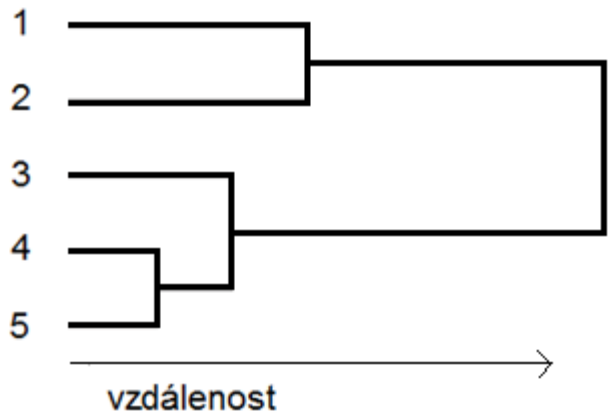
3.3.1 Hierarchické metody

V případě hierarchické shlukové analýzy se jedná o posloupnost rozkladů. Rozlišujeme aglomerativní algoritmus a divizivní algoritmus. **Aglomerativní** přístup je založen na analýze podobnosti. Objekty se spojují podle podobnosti od nejvíce podobných po nejméně podobné. První shluk spojuje tedy pouze dva objekty. Výstupem je jeden shluk. **Divizivní** přístup se zabývá analýzou nepodobnosti. Jeden shluk je postupně rozdělován a výstupem jsou objekty tvořící samostatné shluky. [8]

Hierarchickou shlukovou analýzu bychom dále mohli rozdělit podle počtu proměnných, se kterými počítáme v procesu shlukování. **Monotetický** přístup je speciálním případem divizivního shlukování, kdy je algoritmus použit na binárních datech. Shluky se na určité úrovni vytváří vždy podle jedné z proměnných. **Polytetické** shlukování bere v úvahu všechny proměnné zároveň.

Známe několik **metod** hierarchického shlukování pro určení vzdáleností mezi skupinami ((ne)podobnosti shluků). Mezi nejvýznamnější patří metoda nejbližšího souseda, kdy je vzdálenost skupin je dána minimální vzdáleností objektů. [8] U metody nejvzdálenějšího souseda je sledována naopak maximální vzdálenost objektů. [4] Pro centroidní metodu je určující euklidovská vzdálenost mezi centroidy. Centroidem rozumíme vektory aritmetických průměrů, které byly spočítány pro všechny objekty ve skupině. Dalšími významnými metodami jsou metoda průměrné nepodobnosti objektů, mediánova metoda nebo Wardova metoda. Zde se počítá součet čtverců odchylek jednotlivých hodnot od průměru skupiny. Spojují se skupiny, u nichž je přírůstek tohoto vnitroskupinového součtu minimální. [8]

Výsledky shlukování můžeme zobrazit pomocí hierarchického stromu, tedy stromového diagramu označovaného jako **dendrogram**. Názorně je zde zobrazeno postupné shlukování objektů a tvorba shluků. V horizontální verzi dendrogramu jsou objekty uváděny na ose Y, tzn. na ose Y je zakresleno n listů. Z nich pak vychází větve. Na ose X je zaznamenáno spojení do jedné větve, přičemž se nejdříve spojí objekty, mezi nimiž je nejmenší vzdálenost (nepodobnost). [8]



Obrázek 3: Ukázka zobrazení shluků pomocí dendrogramu

Zdroj: vlastní zpracování

3.3.2 Nehierarchické metody

Podstatou nehierarchických metod je přiřazování objektů do předem stanoveného počtu disjunktních shluků. V případě metody k – průměrů, k – medoidů, k – modů nebo k - histogramů je přiřazování jednotlivých objektů jednoznačné. Můžeme tedy mluvit o metodách pevného shlukování, kdy lze jednoznačně určit, zda objekt ve shluku je nebo není. Naopak v případě fuzzy shlukové analýzy je přiřazení objektů provedeno na základě stanovení míry příslušnosti jednotlivých objektů ke každému ze shluků. [8]

Nehierarchické metody shlukování rozloží množinu do podmnožin podle předem zvoleného kritéria. [9] Kritérium (funkcionál) kvality rozkladu vyjadřuje buď vzájemnou podobnost objektů uvnitř shluků, míru separace shluků, rovnoměrnost rozložení objektů do shluků nebo homogenitu rozložení objektů uvnitř shluků. Provedený rozklad považujeme za optimální, je – li kritérium kvality rozkladu extrémní hodnoty. Funkcionál kvality rozkladu je z matematického pohledu součet čtverců odchylek. [4]

3.3.2.1. Stanovení počtu shluků

Analytik předem obvykle nemá žádnou informaci o počtu shluků. Předpokladem pro použití nehierarchických metod je zadání počtu shluků. Při dobré znalosti dat lze provést expertní odhad. Pro určení počtu shluků můžeme postupovat různými způsoby. Při použití hierarchického shlukování lze stanovit počet shluků na základě dendrogramu (viz oddíl 3.3.1), kde sledujeme velký „skok“ (vzdálenost) mezi shluky. [8] Pro první odhad počtu shluků můžeme využít metodu Kohonenových sítí. Odhadnutý počet shluků většinou představuje horní hranici. Poslední jmenovaný způsob je použit v ukázkovém příkladu v praktické části.

3.3.2.2. Dělení nehierarchických metod

Nehierarchické metody dělíme na optimalizační a analýzu modů. Cílem **optimalizačních metod** je najít optimální rozklad přeřazováním objektů ze shluku do shluku. Kritérium rozkladu se tak minimalizuje nebo maximalizuje. Metoda **analýza modů** chápe shluky jako místa se zvýšenou koncentrací objektů v m – rozměrném prostoru proměnných. [9]

Podstatou **metod zachovávajících stanovený počet segmentů** je výpočet typických (vzorových) bodů existujících skupin objektů. Následuje přiřazení každého objektu k takovému typickému bodu, ke kterému má nejblíže. Tento proces opakujeme do té doby, dokud shluky nebudou stabilní, tj. nebude docházet k přeřazování objektů k jiným typickým bodům. Do této skupiny patří například Forgyova a Janceyova metoda [4] nebo metoda K - means (česky k – průměrů, k – centroidů, MacQueenova metoda), která bude charakterizována v oddílu 3.3.2.3. [8]

Metody s proměnným počtem segmentů umožňují slučování a rozdělování skupin objektů během segmentace. Výsledný počet segmentů se nemusí shodovat s počtem skupin objektů stanovených na počátku shlukování. Na tomto principu je založena například MacQueenova metoda se dvěma parametry, Wishartova metoda RELOC nebo metoda ISODATA. [4]

3.3.2.3. Algoritmus K – means (metoda k - průměrů)

Metodu K – means lze použít pouze v případě, že pracujeme s kvantitativními proměnnými (diskrétními nebo spojitými). Jedná se o optimalizační metodu, která zachová stanovený počet segmentů. Analytik nejprve zadá *k* shluků, do kterých budou objekty přiřazovány. Dále je nutné zadat *k* počátečních typických bodů. Typickými body jsou středy (těžiště) jednotlivých shluků, která jsou někdy označovány jako centroidy. Střed shluku vypočítáme jako průměr všech objektů v jednotlivých shlucích. Počáteční výběr můžeme provést například náhodně nebo výběrem prvních *k* bodů z našeho datového souboru. [9] Vzdálenost každého objektu od každého středu se počítá pomocí euklidovské vzdálenosti (viz oddíl 3.2). Objekt je přiřazen ke středu, ke kterému má nejblíže. [8] Po přiřazení všech objektů do jednotlivých shluků následuje přepočítání středů shluků. [9] Novým středem shluku je m – rozměrný vektor průměrných hodnot jednotlivých proměnných. Po přepočítání středů opět následuje počítání vzdáleností objektů od středů shluků. Opět následuje přiřazování objektů. Pokud má objekt blíže ke středu jiného shluku, je přiřazen do tohoto shluku. [8] Pokud se ve dvou po

sobě jdoucích iteracích nemění rozdělení objektů ve shlucích, shlukování můžeme ukončit. [9]

Algoritmus zahrnuje kroky:

1. Vyber počet shluků *k.*
2. Zadej či vygeneruj *k* počátečních typických bodů.
3. Přiřaď každý objekt ke shluku s nejbližším těžištěm.
4. Přepočítej těžiště.
5. Jestli nedošlo ke změně příslušnosti objektu ke shluku, skonči, jinak pokračuj bodem 3.

Výhodou metody K – means je jednoduchost, rychlost a použitelnost pro velkou množinu dat. Naopak nevýhodou je, že výstupy jsou ovlivněné počátečním výběrem typických bodů a původním přiřazením objektů k těmto těžištím. Výsledky mohou být ovlivněny nežádoucím způsobem i odlehlými objekty. [9]

3.3.2.4. Modifikace K - means

Úpravou metody K – means vznikla metoda k – medoidů, k – modů a k – histogramů. V případě metody **k – medoidů** (PAM – Partitioning Around Medoids) je pro každý vytvořený shluk zjištěn medoid (konkrétní vybraný objekt). Počáteční medoid je stanoven tak, aby byl součet vzdáleností objektů ve shluku od vybraného objektu (medoidu) minimální. [8]

Pro shlukování objektů popsaných pomocí nominálních proměnných se používá metoda k – modů nebo k – histogramů. V případě první uvedené metody je každý shluk charakterizován m – rozměrných vektorem údajů, který obsahuje nejčastěji zastoupené (modální) kategorie proměnných. V případě k – histogramů pracujeme s údaji o četnostech kategorií jednotlivých proměnných. [8]

3.4 Příprava datového souboru

O předzpracování dat bylo již stručně pojednáno v podkapitole 2.3. Fáze přípravy dat je časově nejnáročnější etapou a je velmi důležitá pro úspěšné zpracování úlohy. Provádíme operace s daty jako je například selekce (pro výběr nejdůležitějších atributů z existujících), čištění, vytváření atributů a objektů, integrování a formátování dat. Vytváření nových (odvozování) atributů někdy vyplývá ze známých skutečností, kdy například z rodného čísla odvodíme pohlaví a věk.

3.4.1 Výběr relevantních proměnných

V etapě předzpracování dat by měla být značná pozornost věnována posouzení, zda mají být do analýzy zahrnuty všechny proměnné nebo jen výběr některých. [8] Měli bychom mít na paměti, že data by měla obsahovat pro náš řešený problém relevantní údaje. [1]

Při výběru proměnných je třeba zohlednit statistickou nezávislost proměnných. V souboru se ponechávají proměnné, mezi nimiž není silná závislost. Pro hodnocení závislosti používáme různé koeficienty. Koeficienty závislosti nabývají hodnot z intervalu <0;1>, popř. <-1;1>. Hodnotu nula můžeme interpretovat jako nezávislost. Je nezbytné též určit odlehlé objekty. Po jejich identifikaci je nutné tyto objekty ve vstupní matici vynechat. [8]

3.4.2 Transformace dat

Různé postupy **transformace dat** jsou navrženy pro různé proměnné, např. nominální, ordinální, a dokonce pro data kvantitativní. [8] Pokud jsou hodnoty znaků v různých jednotkách (například q, mg, kalorie), je vhodné před samotným shlukováním provést **standardizaci**, protože metody shlukování by neměly operovat s daty, které jsou závislé na jednotkách měření. [4]

Objekty, které podrobíme shlukové analýze jsou určeny vektory o p složkách, které představují hodnoty p znaků. Normy vektorů mohou nežádoucím způsobem ovlivnit výsledky kvantitativního hodnocení podobností objektů. Někdy je tedy vhodné vektory **normalizovat**, aby měly stejnou normu (nejlépe jednotkovou). [4]

3.4.3 Identifikace odlehlých objektů a práce s chybějícími hodnotami

Je nezbytné též určit odlehlé objekty. Po jejich identifikaci je nutné tyto objekty ve vstupní matici vynechat. [8] Hodnoty výrazně odlišné od ostatních můžeme identifikovat v rámci popisné statistiky například porovnáním aritmetického průměru s useknutým průměrem. Variační koeficient vypovídá o stejnorodosti souboru dat. Extrémy lze snadno vyčíst pomocí histogramu.

Pokud zjistíme, že některý údaj chybí, máme několik možností, jak s danou skutečností pracovat. Objekt s chybějící hodnotou můžeme vynechat nebo chybějící hodnotu nahradit. Nahrazení hodnoty můžeme provést prostřednictvím

- výběrové míry polohy dané proměnné (nejčastěji aritmetický průměr nebo medián),
- hodnoty, která se vyskytuje u jiného objektu, pokud se hodnoty vybraných proměnných shodují,

- podmíněné (skupinové) míry polohy. (Dle vybrané proměnné jsou vytvořeny skupiny. Hodnoty jsou dopočítány z hodnot příslušné skupiny.) [8]

Pro doplnění chybějící hodnoty můžeme využít některý analytický algoritmus pro modelování. [1]

3.5 Okruh problémů velkých datových souborů

Při práci s velkým souborem dat může být největším problémem náročná analýza postavená na výpočtu matice vzdáleností pro všechny dvojice objektů. Dalším problémem je skutečnost, že shlukovací algoritmy efektivně fungují do počtu šestnácti proměnných. V případě velkého počtu atributů je tedy někdy třeba snížit rozměr úlohy na základě analýzy hlavních komponent (PCA – Principal Component Analysis). Úskalí práce s velkým množstvím objektů lze vyřešit pomocí metody shlukování podprostorů.

3.6 Interpretace výsledků

Výklad výsledků úzce souvisí se stanoveným počtem shluků. Na výsledek shlukování má vliv výskyt odlehlých (extrémních) hodnot v datovém souboru. Takové objekty vytvoří samostatné shluky. Platí, že v případě většího počtu odlehlých hodnot v souboru dat se s nastavením většího počtu shluků objevují další a další jednoprvkové shluky na výstupu. Tyto skutečnosti je třeba při interpretaci výsledků zohlednit. [8]

4 PŘÍKLAD – SHLUKOVACÍ METODA K - MEANS

Tato kapitola slouží jako praktický návod pro řešení úloh shlukové analýzy pomocí data miningového nástroje Clementine 10.1 od společnosti SPSS. Pro ukázkový příklad bude použita segmentační metoda K - means. Clementine provádí data miningové analýzy za použití metodologie CRISP – DM. Šest fází této metodologie bude sledovat obsahovou náplň popsanou v teoretické části bakalářské práce.

4.1 Porozumění problematice

Pro přiblížení této problematiky čtenáři byla vybrána data z oblasti občanské vybavenosti obcí všech krajů celé České republiky, jejichž počet obyvatel se pohybuje mezi 5000 a 10000 (vč.). Vzhledem ke skutečnosti, že data pochází z oblasti občanské vybavenosti, nabízí se zde široké využití data miningu s možností zjištění řady užitečných informací z hlediska hodnocení rozvoje obcí. Zaměřila jsem se na existenci kulturních, zdravotnických, školních, sociálních a sportovních zařízení. Data byla získána z portálu Regionálních informačních servisů risy.cz. Většina dat pochází z roku 2010. Data z oblasti kultury, sportu se mi podařilo vyhledat z roku 2006 a údaje ze sociální oblasti jsou datované k roku 2009. Ukázka původních dat je uvedena v příloze A.

Cílem analýzy je vytvořit podklady pro rozhodnutí o přidělení dotační pomoci obcím podle skutečné potřeby odpovídající úrovni jejich rozvoje z hlediska občanské vybavenosti a lépe tak kontrolovat tok finančních prostředků.

4.2 Porozumění datům

4.2.1 Vstupní data

Datový soubor pracuje převážně s informacemi o občanské vybavenosti obcí (5000 – 10000 obyvatel). Jedná se o souhrn obcí z celé České republiky, který obsahuje 140 záznamů a 13 atributů (viz Tabulka 1).

Tabulka 1: Datový slovník

Název atributu	Typ	Typ dat v Clementine	Hodnoty	Popis atributu
OBEC	Text	Set	Bakov nad Jizerou - Železný Brod	Název obce
KRAJ	Text	Set	Jihočeský - Zlínský	Kraj, ve kterém se obec nachází
OKRES	Text	Set	Benešov – Žďár nad Sázavou	Okres, ve kterém se obec nachází
OBYVATELE	Číslo	Range	5004 - 9953	Počet obyvatel obce k 31. 12. 2010
KULTURA	Číslo	Range	2 – 26	Počet kulturních zařízení v obci, bližší popis viz níže
MS	Číslo	Range	1 – 5	Počet mateřských školek v obci
ZS	Číslo	Range	0 – 5	Počet základních škol (nižší i vyšší stupeň) v obci
HRISTE	Číslo	Range	0 – 15	Počet hřišť (s provozovatelem nebo spr.) v obci
TELOCVICNY	Číslo	Range	0 – 15	Počet tělocvičen vč. školních s veřejným přístupem v obci
SPORTOV	Číslo	Range	0 – 18	Počet sportovních zařízení v obci mimo hřiště a tělocvičny, bližší popis viz níže
ZDRAVOT	Číslo	Range	2 – 14	Počet zdravotnických zařízení v obci, bližší popis viz níže
STOMATOLOG	Číslo	Range	0 – 9	Počet samostatných ordinací praktického lékaře stomatologa v obci
SENIOR	Číslo	Range	0 – 8	Počet domovů pro seniory nebo domů s pečovatelskou službou v obci

Zdroj: vlastní zpracování

Před samotnou analýzou je třeba vymezit základní pojmy, kterých se daná problematika dotýká.

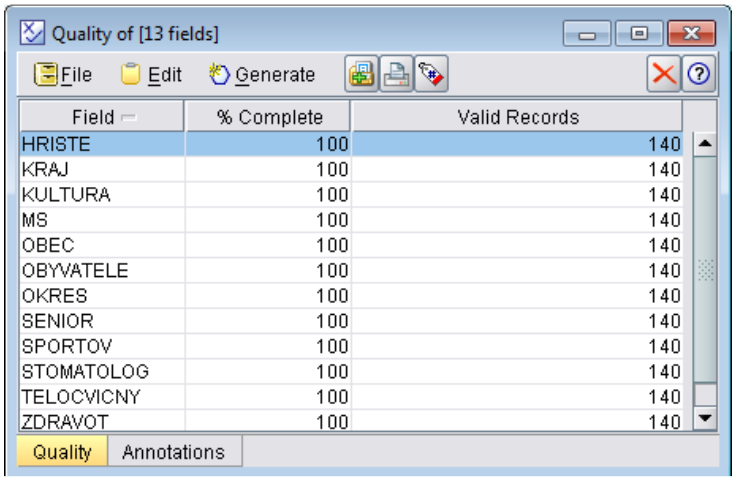
Kulturní zařízení: Zastřešující pojem pro kulturní zařízení typu knihovna (vč. poboček), stálá kina, multikino, divadlo, přírodní amfiteátry (vč. letních kin), muzeum (včetně poboček a samostatných památníků), galerie (vč. poboček a výstavních síní), středisko pro volný čas dětí a mládeže a další kulturní zařízení.

Sportovní zařízení ostatní: Atribut označující koupaliště, kryté bazény, stadiony otevřené, stadiony kryté, zimní stadiony kryté i otevřené a ostatní zařízení pro tělovýchovu. Ostatním sportovním zařízením není myšleno hřiště nebo tělocvičny, které jsou uváděny jako samostatné proměnné.

Zdravotnická zařízení: Proměnná zahrnující sdružená ambulantní zařízení, ambulantní zařízení, nemocnice, detašované pracoviště nemocnic, samostatné ordinace praktického lékaře pro dospělé, samostatné ordinace praktického lékaře pro děti a dorost, střediska záchranné služby a rychlé zdravotnické pomoci, detašované pracoviště středisek záchranné služby a rychlé zdravotnické pomoci nebo okresní zdravotní ústavy. Nejsou zde zahrnuta zařízení lékárenské péče (lékárny), detašované pracoviště zařízení lékárenské péče, léčebny pro dlouhodobě nemocné, samostatné ordinace praktického lékaře gynekologa a samostatné ordinace lékaře specialisty.

4.2.2 Úplnost dat

Pomocí uzlu Quality lze ověřit, zda jsou záznamy v datovém souboru kompletní (Obrázek 4).

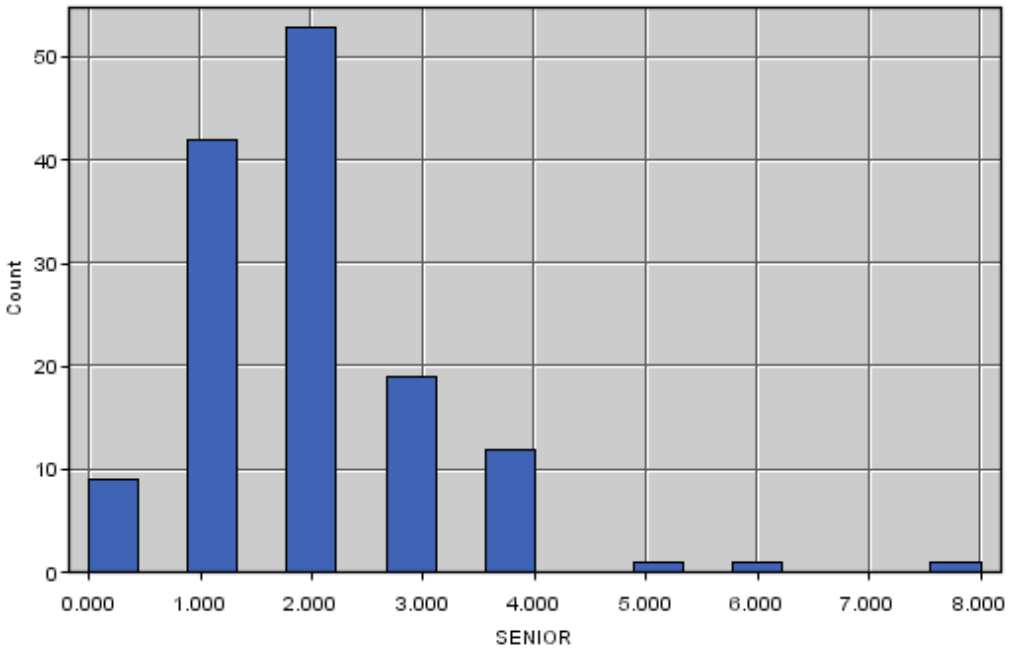


Obrázek 4: Použití uzlu Quality

Zdroj: vlastní zpracování

4.2.3 Popisná statistika, odlehlé hodnoty

V softwarovém prostředí Clementine je pro úvodní analýzu dat proveden datový audit pomocí uzlu Data Audit a statistika pomocí uzlu Statistics. Popisná statistika souboru je zpracována též v prostředí MS Excel 2007 v souboru s názvem „Data“ na listu „Popisná statistika“. Pro jednotlivé proměnné jsou spočítány následující statistické charakteristiky: modus, medián, dolní a horní kvartil, minimum a maximum, rozpětí, aritmetický a geometrický průměr, směrodatná odchylka, variační koeficient a useknutý průměr. Výskyt odlehlých hodnot zvětšuje variabilitu souboru. Odlehlé hodnoty lze odhalit porovnáním useknutého a aritmetického průměru. Velký rozdíl mezi nimi by znamenal přítomnost extrémních hodnot v souboru. V datové sadě se od sebe výrazněji neliší u žádné proměnné. Odlehlé hodnoty lze odhalit i z histogramů, které jsou součástí datového auditu. Ze souboru jsem vyloučila dva záznamy, obec Slavičín a Litovel. V obci Slavičín je osm zařízení pro seniory (domy s pečovatelskou službou nebo domovy pro seniory), což je výrazně více než v ostatních obcích. Obrázek 5 zobrazuje rozdělení četnosti této proměnné. Ze souboru dále vypadla obec Litovel, ve které se nachází dvacet šest kulturních zařízení.



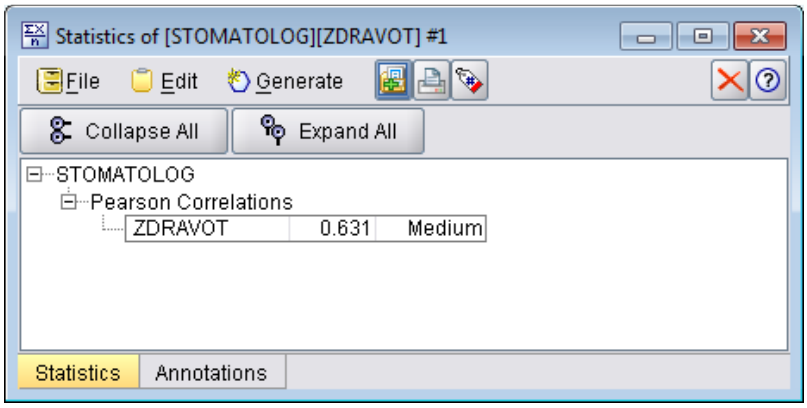
Obrázek 5: Histogram – rozdělení počtu domovů pro seniory a domů s pečovatelskou službou

Zdroj: vlastní zpracování

Nová statistická analýza napovídá, že soubor již neobsahuje extrémní hodnoty.

4.2.4 Korelace

Pomocí uzlu Statistics jsem spočítala korelace mezi spojitými proměnnými. Mezi vlastnostmi neexistuje žádná silná přímá ani nepřímá závislost. (Silná přímá závislost mezi vlastnostmi by byla popsána hodnotami většími než 0,8 a silná nepřímá hodnotami menšími než -0,8.) Středně silnou závislost (0,3 - 0,8), která patří mezi významnější, pozorujeme mezi počtem zdravotnických zařízení a počtem stomatologických zařízení v obci (Obrázek 6).



Obrázek 6: Korelace mezi počtem zdravotnických zařízení a počtem stomatologických zařízení v obci

Zdroj: vlastní zpracování

4.3 Příprava dat

4.3.1 Výběr proměnných

Z třinácti atributů jsem vybrala deset, které jsem považovala pro moji úlohu relevantní. Použila jsem uzel Filter.

4.3.2 Odvození nových proměnných

Vlastnosti, jejichž hodnoty jsou v absolutních číslech, nezohledňují počet obyvatel dané obce. Data, která byla zahrnuta do analýzy, byla převedena na porovnatelné jednotky přepočtem hodnot na 100 obyvatel. Výskyt školních, zdravotních, kulturních, sportovních a sociálních zařízení se tak stal nezávislý na počtu obyvatel obce. Transformace dat byla provedena v MS Excel 2007.

4.3.3 Normalizace dat

Jak bylo uvedeno v oddílu 3.4.2., objekty pro shlukovou analýzu jsou určeny vektory o p složkách, které představují hodnoty p znaků. Protože normy vektorů mohou nežádoucím

směrem ovlivnit výsledky hodnocení podobnosti objektů, byla provedena normalizace. Složky každého z vektorů jsou vyděleny normou tohoto vektoru. Vektory (řádky matice dat) tak nabyly stejnou (jednotkovou) normu. [4] Normalizovaný soubor je součástí excelovského souboru s názvem „Data“. Z normalizovaných dat budeme vycházet ve fázi modelování.

Pomocí uzlu Data Audit jsem provedla novou statistickou analýzu. Popisná statistika souboru normalizovaných dat je součástí přílohy B.

4.4 Modelování

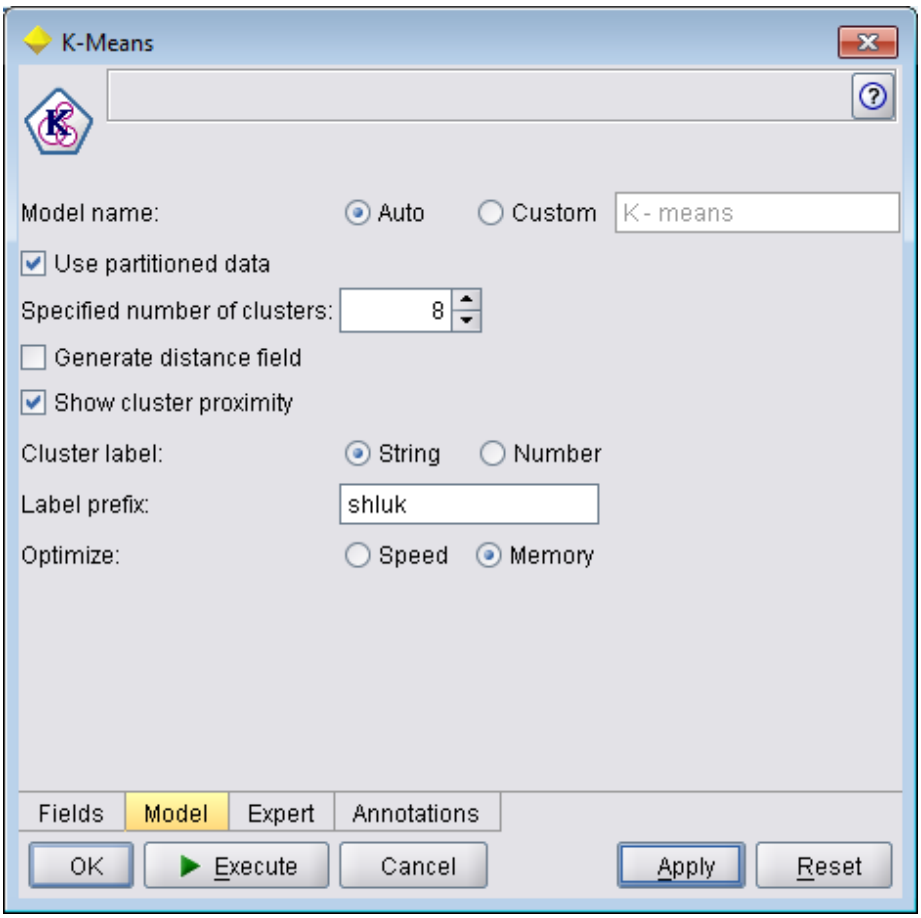
Softwarový nástroj Clementine nabízí uživateli tři seskupovací algoritmy, Kohonenovy mapy, metodu K – means a Two Step. Pro ukázkový příklad byla vybrána metoda K – means.

4.4.1 Metoda K – means

Algoritmus K – means je vhodný je pro analýzu velkého množství dat. Po odstranění odlehlých hodnot obsahuje datový soubor 138 záznamů. Při použití shlukové analýzy na datovém souboru je zapotřebí mít základní představu o analyzovaných vlastnostech.

4.4.2 Nastavení uzlu K – means

Nastavení parametrů provedeme na záložce s názvem **Model** (Obrázek 7). Je důležité rozhodnout se, do kolika shluků bude datový soubor rozdělen (dále viz 4.4.2.1). Počet shluků lze nastavit v poli Specified number of clusters. Pomocí políčka Show cluster proximity zobrazíme informaci o vzájemné vzdálenosti středů vzniklých shluků. Čím je hodnota vyšší, tím jsou si shluky méně blízké. Dále lze pomocí přepínacího tlačítka nastavit formát, v jakém bude zobrazena příslušnost ke shluku. Shluk může být uveden jako číslo (Number) nebo jako řetězec (String). V poli Label prefix je možné shlukům přiřadit specifické pojmenování (například „shluk 1“, „shluk 2“ atd). [2]

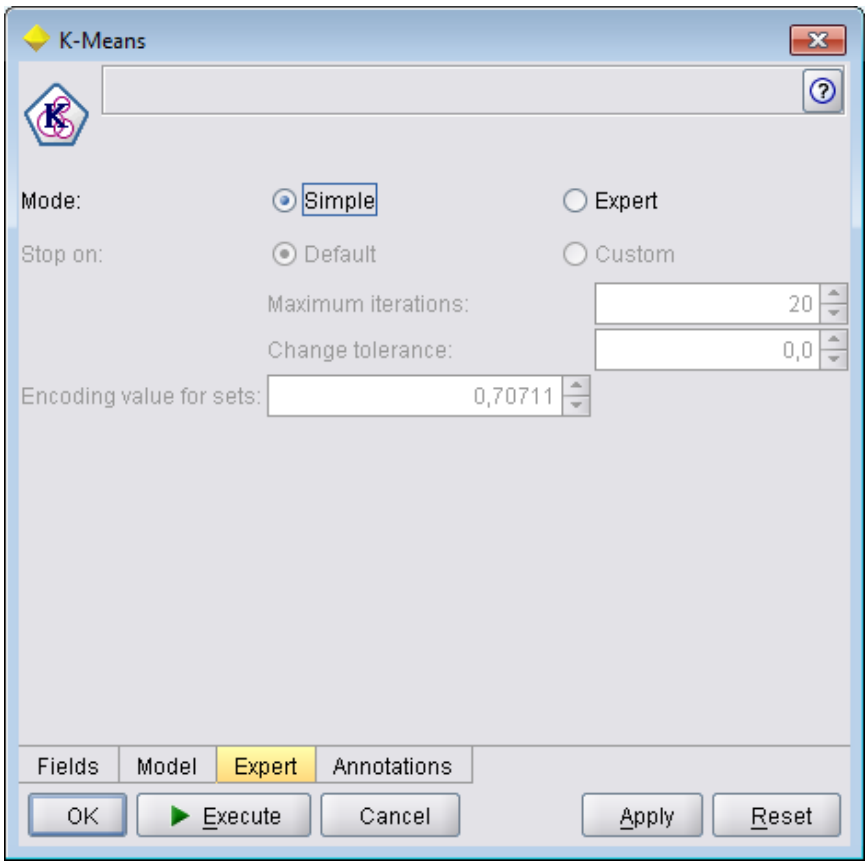


Obrázek 7: Karta Model uzlu K – means

Zdroj: vlastní zpracování

Analytici s odbornými znalostmi metody K – means, ocení kartu **Expert** (Obrázek 8). V režimu Expert na poli s názvem **Maximum iteration** lze omezit počet iterací (přiřazování objektů k typickým bodům). Výchozí hodnota je nastavena na dvacet. Volbou Custom můžeme určit vlastní maximální počet iterací. Hodnota pole s názvem **Chance tolerance** vypovídá o změně příslušnosti objektů ke středům shluků v dané iteraci. Nehierarchické shlukování je zastaveno buď omezením počtu iterací (Maximum iteration) nebo v případě, že míra změny příslušnosti objektů ke středům shluků je v dané iteraci menší než námi specifikovaná hodnota (Chance tolerance). [2]

Encoding value for sets slouží k přiřazení vah kategorizovaným proměnným. Výchozí hodnotou je odmocnina z čísla 0,5 (přibližně 0.707107). Lze zde zadávat hodnoty od nuly do jedničky. Hodnota blíže jedničce přiřadí kategorizované proměnné větší důležitost (váhu) než spojité proměnné. [2]



Obrázek 8: Karta Expert uzlu K – means

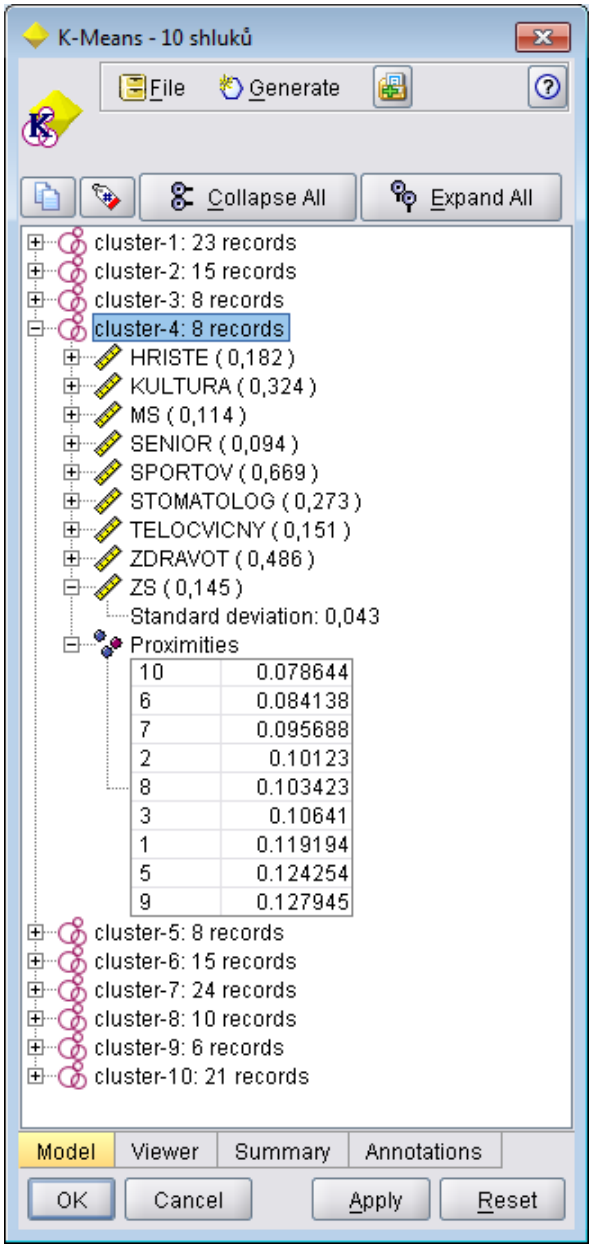
Zdroj: vlastní zpracování

4.4.2.1. Určení počtu shluků

Před rozdělením záznamů do skupin je nutné stanovit počet segmentů. Pro úvodní odhad počtu shluků lze mimo jiné využít segmentační metodu Kohonenovy mapy. Jedná se o speciální druh neuronové sítě, která nevyžaduje přítomnost učitele. [11] Klastr je zde určen dvěma souřadnicemi. Výsledek bude brán jako první odhad počtu shluků, který pak bude využit v metodě K – means. Vstupy pro Kohonenovy mapy jsou již výše zmíněné atributy z oblasti občanské vybavenosti obcí. Vycházíme přitom z normalizovaných dat. Výstup Kohonenových map, dvanáct shluků, zadáme do uzlu K – means. Protože však odhadnutý počet shluků představuje horní hranici, je nutné do uzlu K – means dále zadávat postupně počet shluků vždy o jedna menší a porovnat vzdálenosti mezi shluky (volba proximities). Lepší rozlišení jednotlivých shluků charakterizuje větší vzdálenost (proximities) mezi shluky. Prostřednictvím Kohonenových map, porovnáváním vzdáleností mezi shluky a na základě logické úvahy jsem nastavila konečný počet shluků na deset.

4.4.2.2. Výstup shlukování

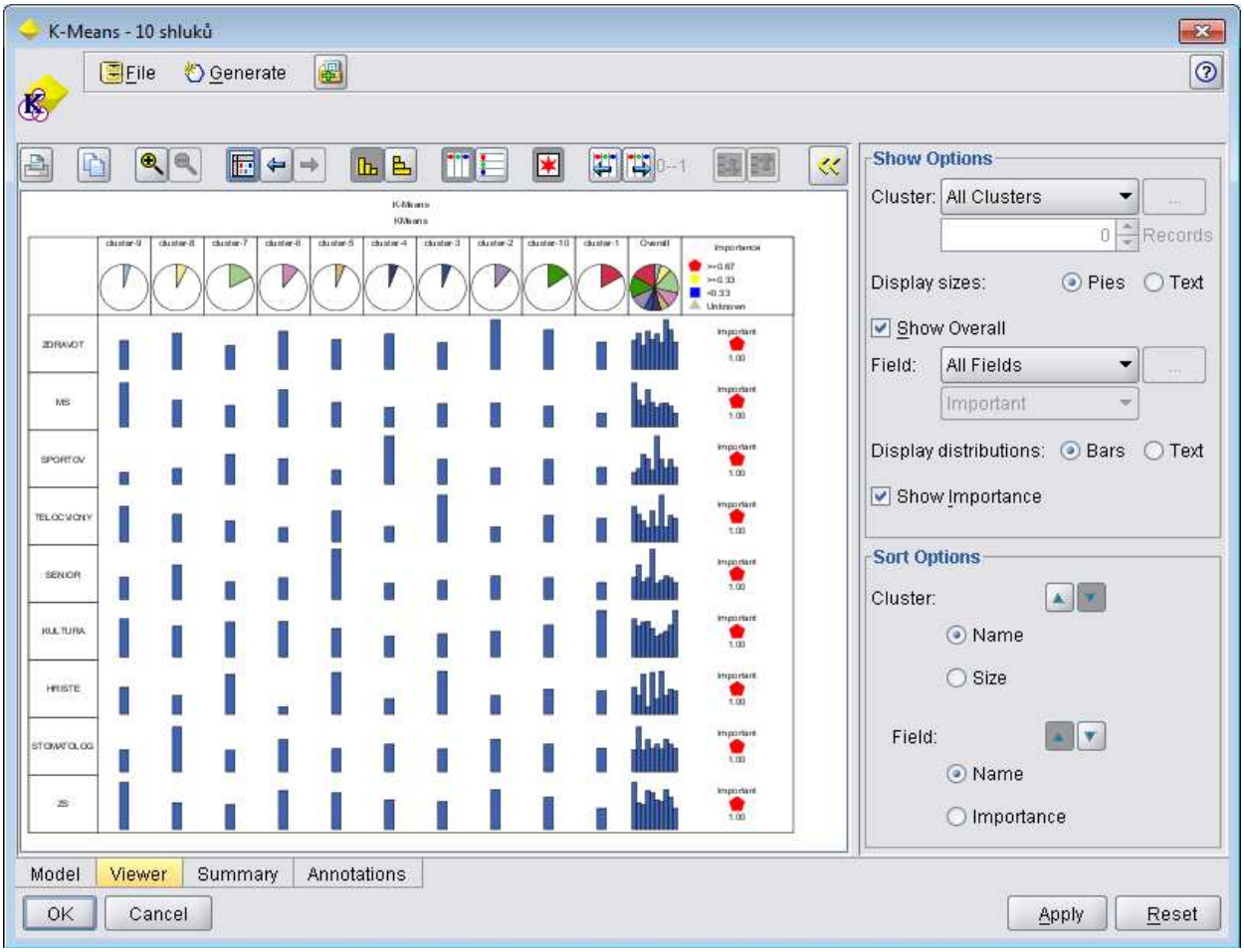
Vstupy pro metodu K – means zůstaly stejné jako v případě Kohonenových map. Pro zpracování metodou K – means nastavím předpokládaný počet shluků na deset. Výstupem je tedy požadovaný počet shluků. Ukázka vzniklých skupin obcí je součástí přílohy C. Obrázek 9 zobrazuje kartu **Model** uzlu K – means (výstup - žlutá ikona), kde je vidět počet záznamů, které náleží jednotlivým shlukům. Na obrázku je zobrazený čtvrtý shluk, který obsahuje osm záznamů. Volba Proximities znázorňuje tabulku se vzdálenostmi mezi shluky. Z obrázku je patrné, že shluk 4 má svými vlastnostmi nejblíže ke shluku 10 (hodnota 0,078644) a nejdál ke shluku 9 (hodnota 0,127945). V závorce u názvu proměnné Počet základních škol (ZS) je uveden její aritmetický průměr (0,145). Směrodatná odchylka (Standard deviation) stejné proměnné dosahuje celkem nízkých hodnot (0,043). Z toho lze usoudit, že se data v daném shluku hromadí okolo průměru.



Obrázek 9: Shluky vygenerované pomocí segmentační metody K – means

Zdroj: vlastní zpracování

Rozdělení shluků do jednotlivých klastrů spolu se zkoumanými atributy je vidět v grafické podobě na kartě **Viewer** (Obrázek 10).



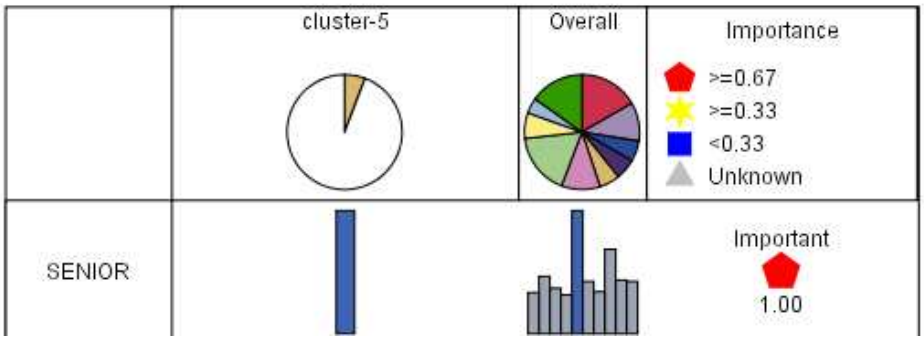
Obrázek 10: Shluky vygenerované metodou K – means a popis vlastností

Zdroj: vlastní zpracování

Ve výchozím nastavení jsou názvy proměnných zobrazeny v levém sloupci a jednotlivé shluky znázorněny jako koláčové grafy v prvním řádku matice. Na rozšířeném okně (žluté šipky) najdeme různé alternativní zobrazení shluků a proměnných. Ve výchozím nastavení jsou shluky uváděné v pořadí sestupně podle počtu záznamů, které obsahují. Pořadí shluků i proměnných lze měnit v oddílu **Sort Options**. Přepínací tlačítka Pies a Text (**Display Size**) udávají styl zobrazení shluků (koláčové grafy nebo procentuální podíl na celkovém počtu obcí). Význam jednotlivých proměnných pro vytváření shluků uvádí sloupec s názvem **Importance**. Z Obrázku 10 je patrné, že všechny naše proměnné jsou pro generování shluků důležité (červený pětiúhelník). Žlutá hvězdička by označovala proměnnou okrajového významu a modrý čtverec proměnnou, která měla jen nepatrný vliv na shlukování. Sloupec podávající celkový přehled přidáme zaškrtnutím políčka **Show Overall** v oddílu Show Options. Přehledový sloupec tvoří skupinu modrých panelů. Znázorňuje rozdělení proměnné v rámci jednotlivých shluků (bars) a umožňuje tak srovnání skupin obcí. Modrý panel představuje střední hodnotu proměnné každého klastru. Nejedná se o histogram zobrazující

četnost rozdělení! Styl zobrazení měníme pomocí přepínacích tlačítek v Display Distributions. Pomocí tlačítka Bars lze vyjádřit atribut jako modrý sloupec. Tlačítko Text znázorní střední hodnotou proměnné pro daný shluk a v závorce směrodatnou odchylku. Tyto statistické charakteristicky lze zobrazit i dvojitým kliknutím na název atributu. [2]

Podrobné informace o jednotlivých shlucích získáme dvojitým kliknutím na schématický koláčových graf jednotlivých shluků nebo výběrem určitého shluku v rozbalovacím seznamu **Cluster** v oddílu Show Options v rozšířeném dialogovém okně. Na Obrázku 11 je vidět, že dominantním rysem shluku 5 je vysoký počet sociálních zařízení v obci (domovy pro seniory, domy s pečovatelskou službou) ve srovnání s ostatními skupinami obcí.



Obrázek 11: Srovnání shluku 6 s celkovým přehledem

Zdroj: vlastní zpracování

V rozbalovacím seznamu Clusters dále můžeme nastavit zobrazení shluků, které obsahují více (méně) záznamů než námi zadaný počet.

Podobně můžeme v rozbalovacím seznamu **Field** nastavit zobrazení jednotlivých proměnných, případně znázornit proměnné specifické důležitosti (Importance) nebo určitého typu (randges, discrete). [2]

Panel nástrojů v uzlu K – means je celkem intuitivní (Obrázek 12). Lze zde například měnit styl zobrazení nebo pomocí červené hvězdičky nastavit hranice významu proměnné (Importance). [2]

Obrázek 12: nástrojů uzlu K – means (žlutá ikona)

Zdroj: vlastní zpracování

4.4.2.3. Vizualizace výsledků

Pro lepší představu o získaných výsledcích nabízí Clementine možnost znázornění výstupů v grafické podobě. Uzel plot napojíme na výstup uzlu K – means (žlutá ikona). Na kartě **Plot**

uzlu Plot vybereme z rozbalovacího seznamu nezávislou proměnnou na osu *x* (X field) a závislou proměnnou na osu *y* (Y field). V oddílu **Overlay** lze nastavit vzhled grafu.

4.4.2.4. Porovnání výsledků shlukovacích metod

K porovnání výsledků více shlukovacích metod lze použít uzel Matrix. Lze zde sledovat shodu ve vygenerovaných záznamech různými shlukovacími metodami.

4.4.2.5. Popis shluků

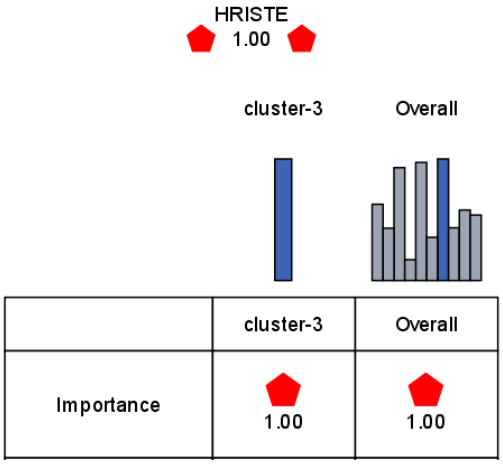
První skupinu tvoří 23 obcí. Pro tuto skupinu je typická dobrá vybavenost z hlediska počtu kulturních zařízení, ale ve srovnání s ostatními obcemi je zde menší počet školních, zdravotnických a sociálních zařízení. Počet obyvatel obcí pohledem na původní data je nevyrovnaný. Do shluku patří obce krajů Jihomoravský, Zlínský, Jihočeský, Plzeňský, Ústecký, Královéhradecký, Středočeský a Vysočina. Obce náležící prvnímu shluku lze vygenerovat pomocí uzlu **Select** definováním podmínky (název proměnné = "hodnota proměnné") v poli Condition na kartě Settings. Na uzel Select napojíme uzel **Table**, prostřednictvím kterého zobrazíme určený výčet obcí. Obce prvního shluku jsou zobrazeny na Obrázku 13.

Obrázek 13: Obce prvního shluku

Zdroj: vlastní zpracování

Druhý shluk zahrnuje obce se slabým kulturním a sportovním zázemím. Nachází se zde zejména nízký počet tělocvičen. Shluk obcí je nadprůměrný z hlediska počtu zdravotnických a stomatologických zařízení. Patří sem celkem patnáct obcí (např. Odolena voda, Úvaly, Jesenice), z nichž téměř polovina pochází ze Středočeského kraje.

Dominantním rysem **třetího shluku** je dobrá vybavenost z hlediska počtu sportovních zařízení. Při pohledu na původní data jistíme, že nadpoloviční většina obcí tohoto shluku má více než deset hřišť. Význam tohoto atributu (Importance) pro sdružení obcí do jednoho shluku je zřejmý z Obrázku 14.

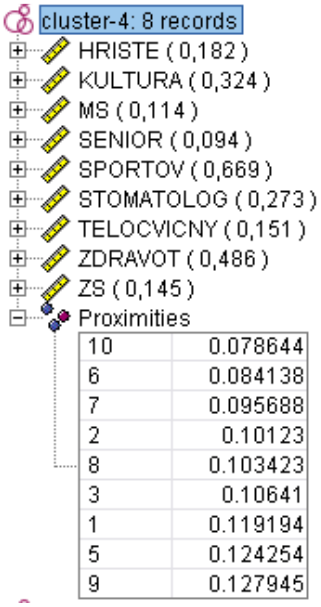


Obrázek 14: Význam atributu počet hřišť pro podobnost záznamů ve čtvrtém shluku

Zdroj: vlastní zpracování

V porovnání s ostatními obcemi je shluk méně rozvinutý v oblasti zdravotnictví. Při pohledu a původní data zjistíme, že je toto uskupení tvořeno osmi obcemi, přičemž šest z nich nepřekračuje hranici 6500 obyvatel. Drtivá většina se nachází na území krajů Moravy.

Čtvrtý shluk vytvořily obce, jejichž výrazným rysem je vynikající zázemí z hlediska ostatních sportovních zařízení (koupaliště a kryté bazény, stadiony otevřené, stadiony kryté, zimní stadiony kryté i otevřené). Velmi dobrou vybavenost vykazuje i v počtu zdravotnických zařízení. Je zde však nízký počet tělocvičen. Z tabulky Proximities (Obrázek 15) lze snadno odvodit, že čtvrtý shluk má nejdále k devátému shluku.



Obrázek 15: Vzdálenost shluků 4 a 9

Zdroj: vlastní zpracování

Výrazným rysem **páté skupiny** je vynikající vybavenost z hlediska počtu sociálních zařízení (domovy pro seniory, domy s pečovatelskou službou). Paradoxně se v obci nachází vysoký počet hřišť. Výrazný podprůměr sledujeme v oblasti kulturních, zdravotnických a ostatních sportovních zařízení (koupaliště a kryté bazény, stadiony otevřené, stadiony kryté, zimní stadiony kryté i otevřené).

Šestý shluk je charakteristický tím, že výrazně převyšuje ostatní obce v počtu mateřských školek. Je zde však nízký počet hřišť a tělocvičen.

Sedmý shluk je typický pro nízkou úroveň v oblasti sociální (domovy pro seniory, domy s pečovatelskou službou). Nejtíživější je zde situace v oblasti zdravotnické a stomatologické péče. Obce sedmé skupiny jsou méně rozvinuté i v oblasti školství. Jediná oblast, ve které převyšuje ostatní obce je sportovní vybavení. Celkově však lze konstatovat, že se jedná o shluk obcí, které patří mezi nejméně vyspělé z pohledu občanské vybavenosti. Tato skupina je s dvaceti čtyřmi obcemi nejpočetnější. Dvacet z nich nepřevyšuje hranici 7000 obyvatel. Výjimku tvoří obce Milovice a Mohelnice, které spadají do největší „velikostní kategorie“. Ze stránky občanské vybavenosti se však rovnají menším obcím.

Obce **osmého shluku** patří výhradně mezi kraje území Čech. Sportovní zázemí obcí osmého shluku je slabé, zejména počet hřišť je silně podprůměrný. Pohybují se však na vyspělé úrovni v oblasti stomatologie.

Devátý shluk tvoří obce s velmi dobrou školní vybaveností, zejména z hlediska počtu mateřských školek. Je zde však i vyšší počet tělocvičen. Nachází se zde však nízký počet zdravotnických, stomatologických a ostatních sportovních zařízení (koupaliště a kryté bazény, stadiony otevřené, stadiony kryté, zimní stadiony kryté i otevřené).

Do **desátého shluku** patří nejrozvinutější obce ve všech zkoumaných atributech. Za povšimnutí stojí zejména vysoký počet zdravotnických zařízení a tělocvičen.

4.5 Vyhodnocení výsledků

Pro zhodnocení rozvoje obcí celé České republiky ve velikostní kategorii od 5000 do 10000 (vč.) jsem použila ukazatele z pěti oblastí:

- Počet kulturních zařízení
- Počet školních zařízení
- Počet sportovních zařízení

- Počet zdravotnických zařízení
- Počet sociálních zařízení

První shluk tvoří obce vyspělé z pohledu dobrého kulturního zázemí. Podobnost obcí druhého seskupení se ukázala v nižším počtu kulturních zařízení a nízkou vybaveností z hlediska sportovního vyžití. Pro třetí shluk je charakteristickým rysem vysoká úroveň v oblasti sportovní vybavenosti. Čtvrtá skupina je typická pro vysoký počet ostatních sportovních zařízení (koupaliště a kryté bazény, stadiony otevřené, stadiony kryté, zimní stadiony kryté i otevřené). Dominantním rysem pátého shluku je vynikající sociální vybavenost (domovy pro seniory, domy s pečovatelskou službou). Do šestého shluku spadají obce s vysokým počtem mateřských školek, ale nízkým počtem hřišť a tělocvičen. Početná sedmá skupina obcí je charakteristická podprůměrnou úrovní z hlediska počtu zdravotnických a stomatologických zařízení. Celkově se jedná o málo rozvinuté obce. Osmý shluk je podprůměrný z hlediska možností sportovního vyžití. Ostatní shluky obcí však převyšuje na poli stomatologie. Devátý shluk je typický pro vysoký počet mateřských školek. Skupina v pořadí desátá zahrnuje celkově rozvinuté obce. Lze zde vysledovat nadprůměr ve všech zkoumaných atributech.

4.6 Využití výsledků

Analýza se zabývala problematikou obcí v České republice ve velikostní kategorii 5000 – 10000 obyvatel (včetně). Rozvoj obcí byl hodnocen z hlediska občanské vybavenosti. Cílem analýzy ukázkového příkladu bylo vytvořit podklad pro rozhodnutí o přidělení dotační pomoci podle skutečné potřeby rozvoje obcí. Podle výsledků analýzy je třeba zaměřit se na obce sedmého shluku, které patří mezi nejméně rozvinuté. Nejproblematičtější situace se ukázala ve zdravotnictví a stomatologii.

Pro případné další analýzy by bylo vhodné soubor doplnit o další atributy (např. ekonomické, demografické, strukturální). V návaznosti na řešený problém je vhodné využít další dataminingové metody (rozhodovací stromy atd.)

ZÁVĚR

V teoretické části byly naznačeny typické úlohy řešené pomocí data miningu. Práce se dále věnuje aplikačním oblastem data miningu a metodám, které zastřešuje tento pojem. Bylo poukázáno na metodiky, které data mining používá a rozebrána metodika CRISP – DM z pohledu náplně jednotlivých fází. Třetí kapitola čtenáře seznámila s metodou shlukové analýzy se zaměřením na nehierarchickou metodu K – means.

V praktické části byl zpracován příklad, který bude podkladem pro studenty předmětu PZDM. Analýza byla provedena pomocí data miningového nástroje Clementine 10.1 od společnosti SPSS. Pro ukázkový příklad byla zvolena datová sada z oblasti občanské vybavenosti obcí České republiky spadajících do velikostní kategorie od 5000 do 10000 obyvatel. Cílem analýzy bylo rozdělit obce pomocí shlukové analýzy do skupin, které jsou si na základě zvolených atributů podobné. Příklad využil zásad metodologie CRISP – DM a sledoval náplň jednotlivých fází popsaných v teoretické části práce.

Nejprve byly stanoveny cíle projektu. Následovalo vypracování datového slovníku, bližší seznámení s daty a zpracování popisné statistiky pro lepší vhled do zkoumaných dat. V rámci přípravy dat následoval výběr relevantních proměnných a datový soubor byl očištěn od odlehlých hodnot. V MS Excel 2007 proběhla normalizace dat. Ve fázi modelování bylo nutné stanovit počet shluků. Pro rozdělení souboru do shluků jsem vybrala nehierarchickou metodu K – means. Po interpretaci výsledků byly výsledky vyhodnoceny a posouzeny z manažerského úhlu pohledu s ohledem na účel úlohy, který byl definován v první fázi data miningového procesu. Výstupem byl tedy podklad pro rozhodnutí o přidělení dotační pomoci obcím. Jako nejproblematičtější se ukázala situace v obcích sedmého shluku.

Bakalářská práce byla koncipována jako průvodce pro čtenáře při zpracování datového souboru pomocí shlukové analýzy v softwarovém prostředí Clementine 10.1 od společnosti SPSS. Její snahou bylo poskytnout studentovi bez větších zkušeností se zmíněnou aplikací jakousi případovou studii, která by přispěla k lepší orientaci v programu. Práce by studenta zároveň měla naučit sledovat správné zásady metodologie CRISP.- DM. Podařilo se popsat práci s odlehlými hodnotami, postup normalizace dat, nastínit postup zpracování příkladu v softwarovém prostředí Clementine.

POUŽITÁ LITERATURA

[1] BERKA, P. *Dobývání znalostí z databází.* Praha: Academia, 2003, 366 s. ISBN 80-200-1062 9.

[2] *Clementine User´s 8.0 Guide* [online]. Integral Solution Limited, 2003 [cit. 2012-04-29]. ISBN 1-568-27-333-9. Dostupné z: http://stat.smmu.edu.cn/DOWNLOAD/ebook/Clementine%20Users%20Guide.pdf

[3] DOSTÁL, P., RAIS, K., SOJKA, Z. *Pokročilé metody manažerského rozhodování: konkrétní příklady využití metod v praxi.* Praha: Grada, 2005, 166 s. ISBN 80–247–1338-1

[4] LUKASOVÁ, A., ŠARMANOVÁ, J. *Metody shlukové analýzy.* Praha: SNTL, 1985, 210 s. ISBN 04 – 014 - 85

[5] PETR, P. *Data mining, díl 1.* Pardubice: Univerzita Pardubice, 2006. 144 s. ISBN 80-7194-886-1.

[6] *Portál Regionálních Informačních Servisů* [online]. [cit. 2012-04-18]. Dostupné z: http://www.risy.cz/cs/krajske-ris/jihomoravsky-kraj/verejna-sprava/spravni-cleneni/obce/

[7] RUD, O. L. *Data Mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM).* Praha: Computer Press, 2001, 329 s. ISBN 80-7226-577-6

[8] ŘEZÁNKOVÁ, H., HÚSEK, D., SNÁŠEL, V. *Shluková analýza dat.* Praha: Professional Publishing, 2009, 220 s. ISBN 978-80-86946-81-8

[9] *Shluková analýza*: *Nehierarchické metody shlukování* [online]. [cit. 2012-04-18]. Dostupné z: http://is.muni.cz/th/172767/fi_b/5739129/web/web/nehiermet.html

[10] SKALSKÁ, H. *Data mining a klasifikační modely.* Hradec Králové: Gaudeamus, 2010, 154 s. ISBN**:** 978-80-7435-088-7

[11] SKLENÁK, V. *Data, informace, znalosti a Internet*. Praha: C H Beck, 2001. 507 s. ISBN 80-7179-409-0

[12] *Step-by-step data mining guide* [online]. 1.0. c2000 [cit. 2011-06-27]. CRISP-DM 1.0. Dostupné z WWW: http://community.udayton.edu/provost/it/training/documents/SPSS_CRISPWPlr.pdf

SEZNAM PŘÍLOH

Příloha A: Ukázka původních dat

KRAJ	OKRES	OBEC	OBYVATELE	KULTURA	MS	ZS	HRISTE	TELOCVICNY	SPORTOV	ZDRAVOT	STOMATOLOG	SENIOR
Jihomoravský	Blansko	Letovice	6892	22	1	1	3	4	7	6	5	2
	Brno - venkov	Ivančice	9376	11	4	5	5	6	4	8	3	0
	Brno - venkov	Rosice	5570	9	2	1	3	3	4	5	3	0
	Brno - venkov	Šlapanice	7021	11	2	1	5	5	5	5	4	1
	Brno - venkov	Tišnov	8662	9	3	2	12	15	10	11	8	2
	Břeclav	Hustopeče	5962	7	2	3	0	3	5	8	5	0
	Břeclav	Mikulov	7450	11	2	3	4	4	3	8	5	2
	Hodonín	Dubňany	6474	3	2	1	4	4	3	3	2	1
	Hodonín	Strážnice	5756	12	1	2	3	3	3	7	4	2
	Vyškov	Bučovice	6455	12	1	2	7	6	1	9	4	2
	Vyškov	Rousínov	5363	7	1	1	6	4	4	6	2	1
	Vyškov	Slavkov u Brna	6245	7	1	3	1	3	4	7	9	6
	Znojmo	Moravský Krumlov	5977	15	1	2	5	6	0	8	5	1
Zlínský	Kroměříž	Bystřice pod Hostýne	8603	10	5	2	7	5	9	13	4	1
	Kroměříž	Hulín	7196	6	1	1	3	3	5	6	3	2
	Kroměříž	Chropyně	5112	9	1	0	4	3	3	4	2	2
	Uherské Hradiště	Kunovice	5504	7	1	2	5	3	3	7	2	1
	Uherské Hradiště	Staré Město	6821	4	3	0	4	6	4	6	8	2
	Vsetín	Zubří	5618	6	2	1	6	2	5	3	2	0
	Vsetín	Brumov-Bylnice	5828	14	1	1	6	2	6	5	3	1
	Vsetín	Luhačovice	5247	12	1	1	4	2	3	5	5	2
	Vsetín	Napajedla	7423	6	1	2	4	3	4	9	6	3
	Vsetín	Slavičín	6800	14	2	3	13	3	4	10	6	8
	Vsetín	Valašské Klobouky	5088	11	1	2	7	4	8	7	6	2

Zdroj: vlastní zpracování

Příloha B: Datový audit normalizovaných dat

Data Audit of [10 fields] #1

File Edit Generate

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
KULTURA		Range	0.097	0.880	0.526	0.145	-0.029	$null$	138
MS		Range	0.036	0.394	0.138	0.077	0.979	$null$	138
ZS		Range	0.000	0.359	0.151	0.065	0.399	$null$	138
HRISTE		Range	0.000	0.616	0.307	0.158	0.087	$null$	138
TELOCVIC...		Range	0.000	0.631	0.233	0.115	0.638	$null$	138
SPORTOV		Range	0.000	0.717	0.319	0.155	0.430	$null$	138
ZDRAVOT		Range	0.144	0.808	0.461	0.132	0.068	$null$	138
STOMATOL...		Range	0.000	0.570	0.270	0.102	0.439	$null$	138
SENIOR		Range	0.000	0.533	0.132	0.087	1.317	$null$	138

Audit Annotations

Zdroj: vlastní zpracování

Příloha C: Ukázka shluků

SHLUK 1	SHLUK 2	SHLUK 3	SHLUK 4	SHLUK 5	SHLUK 6	SHLUK 7	SHLUK 8	SHLUK 9	SHLUK 10
Letovice	Napajedla	Tišnov	Bystřice	Petrovice u Karviné	Hustopeče	Rousínov	Slavkov u Brna	Ivančice	Bystřice pod Hostýnem
Rosice	Jablunkov	Dubňany	Frýdlant nad Ostravicí	Petřvald	Polička	Zubří	Staré Město	Sezimovo Ústí	Hulín
Šlapanice	Ledeč nad Sázavou	Vrbno pod Pradědem	Kdyně	Příbor	Choceň	Valašské Klobouky	Holice	Nejdek	Kunovice
Mikulov	Blatná	Dolní Lutyně	Nýřany	Přelouč	Milevsko	Rychvald	Vimperk	Habartov	Rýmařov
Strážnice	Planá	Vítkov	Lovosice	Horní Slavkov	Kraslice	Bílovec	Bechyně	Červený Kostelec	Odry
Bučovice	Česká Kamenice	Šenov	Tanvald	Mimoň	Šluknov	Fulnek	Veselí nad Lužnicí	Hostivice	Žamberk
Moravský Krumlov	Štětí	Lipník nad Bečvou	Jilemnice	Broumov	Dubí	Hradec nad Moravicí	Přeštice		Moravské Budějovice
Chropyně	Podbořany	Železný Brod	Lomnice nad Popelkou	Hronov	Duchcov	Kravaře	Nová Paka		Náměšť nad Oslavou
Brumov-Bylnice	Český Brod				Chlumec nad Cidlinou	Vratimov	Týnec nad Sázavou		Týn nad Vltavou
Luhačovice	Mnichovo Hradiště				Hořovice	Kojetín	Nové Strašecí		Kaplice
Chotěboř	Lysá nad Labem				Králův Dvůr	Mohelnice			Vodňany
Světlá nad Sázavou	Odolena Voda				Benátky nad Jizerou	Skuteč			Soběslav
Polná	Úvaly				Černošice	Letohrad			Nýrsko
Telč	Jesenice				Roztoky	Velká Bíteš			Stříbro
Třešť	Sedlčany				Dobříš	Dobřany			Frýdlant
Bystřice nad Pernštejnem						Františkovy Lázně			Hrádek nad Nisou
Dačice						Jílové			Chrastava
Třeboň						Doksy			Semily
Horažďovice						Česká Skalice			Nový Bydžov
Postoloprty						Dobruška			Hořice
Třebechovice pod Orebem						Kostelec nad Orlicí			Nové Město nad Metují
Úpice						Týniště nad Orlicí			
Bakov nad Jizerou						Stochov			
						Milovice			

Zdroj: vlastní zpracování