

PŘÍRODOVĚDECKÁ FAKULTA UNIVERZITY PALACKÉHO
KATEDRA INFORMATIKY

DIPLOMOVÁ PRÁCE

Analýza dat ze studentských dotazníků



2013

Bc. Tomáš Matonoha

Anotace

Data mining je proces získávání netriviálních a dříve neznámých informací z dat. Na datech z dotazníků při zápisu studentů do 1. ročníku Přírodovědecké fakulty UP v Olomouci z let 2010 a 2011 jsem provedl dolování dat pomocí asociačních pravidel, hierarchického shlukování a metody GUHA. Výsledkem mé práce jsou analyzované shluky a zajímavá pravidla, která se dají využít pro další účely.

Děkuji doc. RNDr. Michalu Krupkovi, Ph.D. za vedení této diplomové práce
a za rady při konzultacích.

Obsah

1. Úvod	10
2. Dobývání znalostí z databází	11
2.1. Metodika CRISP-DM	11
2.2. Asociační pravidla	13
2.2.1. Algoritmy hledání asociačních pravidel	14
2.3. Metoda GUHA	16
2.3.1. Pojmy	16
2.3.2. Kvantifikátory	17
2.3.3. Procedury metody GUHA	18
2.4. Shluková analýza	18
2.4.1. Hierarchické shlukování	20
2.4.2. Nehierarchické shlukování	22
3. Software	23
3.1. LISp-Miner	23
3.1.1. Struktura systému	23
3.1.2. Procedura 4ft-Miner	24
3.2. R	26
3.2.1. Asociační pravidla	27
3.2.2. Shlukování	27
3.2.3. k-means	28
4. Data mining dotazníků při zápisu do 1. ročníku PřF UP	28
4.1. Struktura dotazníků	29
4.2. Příprava dat	30
4.3. Rok 2011 - Apl. informatika a Informatika	30
4.3.1. Asociační pravidla	30
4.3.2. Hierarchické shlukování	31
4.3.3. k-means	36
4.3.4. Metoda GUHA	37
4.4. Rok 2010 - Apl. informatika a Informatika	42
4.4.1. Hierarchické shlukování	42
4.4.2. Metoda GUHA	44
4.5. Rok 2010 a 2011 - Apl. informatika a Informatika	46
4.5.1. Hierarchické shlukování	46
4.5.2. Metoda GUHA	47
4.6. Rok 2011 - všechny obory	50
4.6.1. Hierarchické shlukování	50
4.6.2. Metoda GUHA	54
4.7. Rok 2010 - všechny obory	55

4.7.1. Hierarchické shlukování	55
4.7.2. Metoda GUHA	60
4.8. Rok 2010 a 2011 - všechny obory	61
4.8.1. Hierarchické shlukování	61
4.8.2. Metoda GUHA	66
4.9. Vyhodnocení výsledků	67
4.9.1. Interpretace pro zákazníka	68
4.10. Závěr	69
Závěr	70
Conclusions	71
Reference	72
A. Obsah přiloženého CD	73
B. Dendrogramy hierarchického shlukování	74
C. Typičtí zástupci shluků	97

Seznam obrázků

1.	Fáze metodiky CRISP-DM	12
2.	Metoda nejbližšího souseda	21
3.	Metoda nejvzdálenějšího souseda	21
4.	Metoda průměrné vzdálenosti	21
5.	Centroidní metoda	22
6.	LISp-Miner - Modul 4ftTask.exe	24
7.	Dendrogram shlukování - Apl. inf. a Informatika 2011 (single linkage)	75
8.	Dendrogram shlukování - Apl. inf. a Inf. 2011 (average linkage) . .	76
9.	Dendrogram shlukování - Apl. inf. a Inf. 2011 (complete linkage) .	77
10.	Dendrogram shlukování - Apl. inf. a Inf. 2011 (Wardova metoda) .	78
11.	Dendrogram shluk. - Apl. inf. a Inf. 2011 - jádra (complete linkage)	79
12.	Dendrogram shluk. - Apl. inf. a Inf. 2011 - jádra (Wardova metoda)	80
13.	Dendrogram shlukování - Apl. inf. a Inf. 2010 (complete linkage) .	81
14.	Dendrogram shlukování - Apl. inf. a Inf. 2010 (Wardova metoda) .	82
15.	Dendrogram shluk. - Apl. inf. a Inf. 2010 a 2011 (complete linkage)	83
16.	Dendrogram shluk. - Apl. inf. a Inf. 2010 a 2011 (Wardova metoda)	84
17.	Dendrogram shlukování - 2011 (complete linkage)	85
18.	Dendrogram shlukování - 2011 (Wardova metoda)	86
19.	Dendrogram shluk. s obory - 2011 (complete linkage)	87
20.	Dendrogram shluk. s obory - 2011 (Wardova metoda)	88
21.	Dendrogram shlukování - 2010 (complete linkage)	89
22.	Dendrogram shlukování - 2010 (Wardova metoda)	90
23.	Dendrogram shluk. s obory - 2010 (complete linkage)	91
24.	Dendrogram shluk. s obory - 2010 (Wardova metoda)	92
25.	Dendrogram shlukování - 2010 a 2011 (complete linkage)	93
26.	Dendrogram shlukování - 2010 a 2011 (Wardova metoda)	94
27.	Dendrogram shluk. s obory - 2010 a 2011 (complete linkage) . . .	95
28.	Dendrogram shluk. s obory - 2010 a 2011 (Wardova metoda) . . .	96

Seznam tabulek

1.	Frekvenční (čtyřpolní) tabulka.	13
2.	Příklad tabulky dat studentů a jejich pozornosti.	16
3.	Vybraná asoc. pravidla (min. podpora 0,1; min. spolehlivost 0,5).	31
4.	Tabulka odpovědí a vzdáleností s jednou opačnou odpovědí.	34
5.	Tabulka odpovědí a vzdáleností se třemi různými odpověďmi.	34
6.	Tabulka odpovědí a vzdáleností se třemi různými odpověďmi.	35
7.	Tabulka odpovědí a vzdáleností se třemi různými odpověďmi.	35
8.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	36
9.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	36
10.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	37
11.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	37
12.	Tabulka proc. podílů shluků jednotlivých vah u complete linkage první kategorie.	38
13.	Tabulka proc. podílů shluků jednotlivých vah u complete linkage druhé kategorie.	38
14.	Tabulka proc. podílů shluků jednotlivých vah u complete linkage třetí kategorie.	39
15.	Tabulka proc. podílů shluků jednotlivých vah u Wardovy metody první kategorie.	39
16.	Tabulka proc. podílů shluků jednotlivých vah u Wardovy metody druhé kategorie.	40
17.	Tabulka proc. podílů shluků jednotlivých vah u Wardovy metody třetí kategorie.	40
18.	Tabulka hodnot podpory a spolehlivosti.	41
19.	Tabulka hodnot podpory a spolehlivosti.	41
20.	Tabulka hodnot podpory a spolehlivosti.	42
21.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	44
22.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	44
23.	Tabulka hodnot podpory a spolehlivosti.	45
24.	Tabulka hodnot podpory a spolehlivosti.	46
25.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	48
26.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	48
27.	Tabulka hodnot podpory a spolehlivosti.	49

28.	Tabulka hodnot podpory a spolehlivosti.	50
29.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	51
30.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	52
31.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	53
32.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	54
33.	Tabulka hodnot podpory a spolehlivosti.	56
34.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	57
35.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	58
36.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	59
37.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	60
38.	Tabulka hodnot podpory a spolehlivosti.	61
39.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	63
40.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	63
41.	Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.	65
42.	Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.	65
43.	Tabulka hodnot podpory a spolehlivosti.	67
44.	Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - complete linkage .	98
45.	Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - Wardova metoda .	99
46.	Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - complete linkage .	100
47.	Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - Wardova metoda .	101
48.	Typičtí zástupci shluků - Apl. inf. a Inf. 2010 - complete linkage .	102
49.	Typičtí zástupci shluků - Apl. inf. a Inf. 2010 - Wardova metoda .	103
50.	Typičtí zástupci shluků - Apl. inf. a Inf. 2010 a 2011 - complete linkage	104
51.	Typičtí zástupci shluků - Apl. inf. a Inf. 2010 a 2011 - Wardova metoda	105
52.	Typičtí zástupci shluků - 2011 - complete linkage	106
53.	Typičtí zástupci shluků - 2011 - Wardova metoda	108
54.	Typičtí zástupci shluků s obory - 2011 - complete linkage	109
55.	Typičtí zástupci shluků s obory - 2011 - Wardova metoda	110
56.	Typičtí zástupci shluků - 2010 - complete linkage	111

57.	Typičtí zástupci shluků - 2010 - Wardova metoda	112
58.	Typičtí zástupci shluků s obory - 2010 - complete linkage	113
59.	Typičtí zástupci shluků s obory - 2010 - Wardova metoda	114
60.	Typičtí zástupci shluků - 2010 a 2011 - complete linkage	115
61.	Typičtí zástupci shluků - 2010 a 2011 - Wardova metoda	116
62.	Typičtí zástupci shluků s obory - 2010 a 2011 - complete linkage .	118
63.	Typičtí zástupci shluků s obory - 2010 a 2011 - Wardova metoda .	119

1. Úvod

Dobývání znalostí z dat je disciplína, jejíž kořeny sahají do 60. let 20. století. Využívá se hojně v komerční oblasti (marketing, telekomunikace, finance) a vědeckém výzkumu. S neustálým zvyšováním rychlosti a paměti počítačů rostl požadavek po zpracování velkého množství dat. Dolování dat je myšleno jako automatický nebo poloautomatický proces. Jeho výsledkem jsou netriviální a dříve neznámé informace nebo závislosti, které je nutné také vhodně formulovat pro další využití.

Cílem této práce je nastudovat a aplikovat vhodné metody dobývání znalostí (především metodu GUHA) na dotazníky při zápisu do 1. ročníku Přírodovědecké fakulty UP z let 2010 a 2011.

2. Dobývání znalostí z databází

Dobývání znalostí z databází, označováno také jako data mining, je proces získávání netriviálních a dříve neznámých informací z dat.

Tento proces je většinou iterativní a má obecně několik kroků, které vedou k určitému cíli. Je to selekce, předzpracování, transformace, dolování dat a interpretace [1] a [2].

Předzpracování dat zahrnuje přípravu dat do formy vyžadované pro aplikaci konkrétních metod. Samotné dolování dat je prováděno analytickými metodami, které mohou být prováděny vícekrát za sebou s různými hodnotami parametrů. Cílem interpretace je vhodné zpracování velkého množství výsledků dolování dat a interpretace cílovým uživatelům.

Při řešení problému je nutné ho nejdříve specifikovat a poté je třeba získat všechna dostupná data pro jeho řešení. Existuje několik metod analýzy dat a často je potřeba tyto metody kombinovat navzájem. Mezi nejznámější metody patří klasifikační metody, asociační pravidla, metody explorační analýzy, rozhodovací stromy, genetické algoritmy, neuronové sítě, hrubé množiny atd.

V průběhu vývoje byly navrženy metodiky, které poskytují jednotné postupy pro řešení úloh při dolování dat. Některé metodiky vznikly ve výzkumných a komerčních institucích či firmách. Mezi nejznámější metodiky patří 5A, SEMMA a CRISP-DM, kterou blíže představím.

Tato kapitola obsahuje teoretický přehled metod dolování dat, které jsem na dotazníky použil.

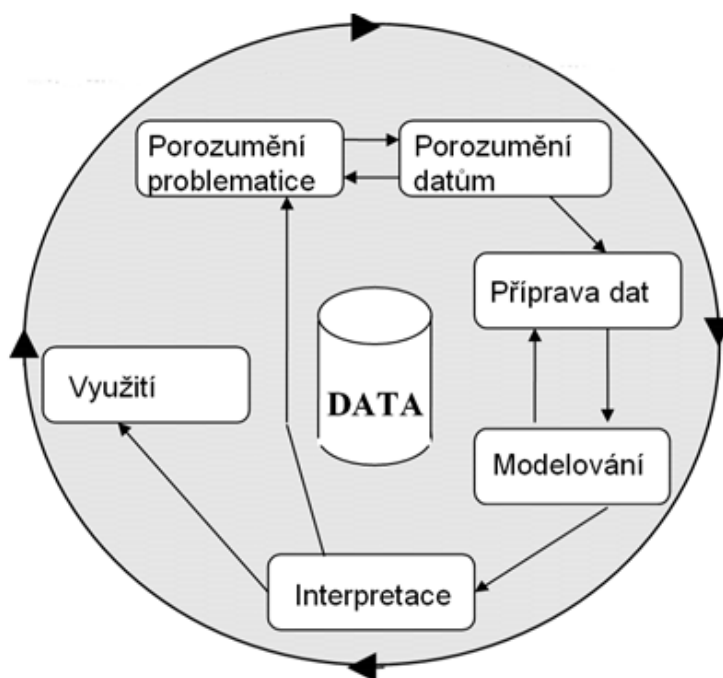
2.1. Metodika CRISP-DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) [3] je jednou z nejpoužívanějších metodik [4]. Byla poprvé publikována v roce 1998 sdružením firem zabývajících se dolováním dat. Jejím cílem je vytvořit univerzální postup při dolování dat a umožnit řešit úlohy rychleji, efektivněji, spolehlivěji a s nižšími náklady.

Myšlenkou této metodiky je znázornění procesu získávání znalostí jako neustále se opakující posloupnosti kroků. Metodiku tvoří celkem šest fází, jejichž pořadí není pevně stanoveno. Výsledek jedné fáze ovlivňuje fáze následující a často je třeba se vrátit k předchozím fázím. Celý tento proces je znázorněn na obrázku 1.

1. **Porozumění problematice** - první krok procesu obsahující pochopení cílů a požadavků. Je nutné porozumět zákazníkovi, co vlastně vyžaduje.
2. **Porozumění datům** - prvotní seznámení a analýza dat dodaných zákazníkem.

3. **Příprava dat** - časově nejnáročnější fáze metodiky. Zahrnuje převedení dat od zákazníka do podoby vhodné pro zpracování dalšími nástroji. V této fázi se vybírají data vhodná pro dolování, zjišťují se chybějící hodnoty, probíhá slučování tabulek, diskretizace atributů a formátování dat [5].
4. **Modelování** - v této fázi se na data iterativně aplikují modelovací techniky a vyladují se optimální hodnoty parametrů. Pro řešení úlohy se použije jedna nebo více dolovacích technik. Pro každou techniku se nastaví vhodné parametry a metriky.
5. **Vyhodnocení výsledků** - výsledky předchozí fáze musejí být před předáním zákazníkovi zhodnoceny. To zahrnuje i aspekt, zda jsme vůbec našli nové informace. Případně je také nutné opakovat jednotlivé fáze nebo začít celé dobývání dat znovu od začátku.
6. **Využití výsledků** - výsledky dolování musejí být zákazníkovi dodány v takovém formátu, aby je mohl dále využít.



Obrázek 1.: Fáze metodiky CRISP-DM

V metodice je kladen velký důraz na přípravnou fázi porozumění problematice a porozumění reprezentaci problematiky v datech. Po této fázi následuje předzpracování dat a analýza.

Tato metodika pokrývá všechny fáze dobývání znalostí z databází a zdůrazňuje jeho cyklický charakter. Podle této metodiky jsem se snažil v mé praktické části diplomové práce řídit.

2.2. Asociační pravidla

Asociační pravidla jsou zajímavé vazby ve velkém množství dat a patří k nejčastějším způsobům vyjadřování znalostí. Samotný termín *asociační pravidlo* zpopularizovala v 90. letech skupina kolem Agrawala, která se zabývala analýzou nákupního košíku (angl. market basket analysis [6]). Tato analýza zjišťuje závislosti mezi zbožím, které si zákazníci vložili do svého nákupního košíku. Znalost této informace se může poté využít pro marketingové účely, rozmístění zboží v prodejně atd.

Všechny následující metody dobývání dat z databází pracují nad daty, které jsou ve formě tabulky viz 2. Tato tabulka neboli matice obsahuje obecně m řádků (záznamů) a n sloupců (atributů).

Asociační pravidlo je tento výraz:

$$Ant \Rightarrow Suc \text{ [podpora} = p, \text{spolehlivost} = s]$$

kde Ant i Suc jsou vzájemně disjunktní množiny konjunkcí hodnot atributů. Ant nazýváme předpoklad a Suc závěr. Výraz vyjadřuje, že záznamy obsahující Ant mají tendenci obsahovat také Suc .

Základními charakteristikami asociačních pravidel jsou podpora (*support*) a spolehlivost (*confidence*). Asociace, které překročí uživatelem zadané minimální hodnoty podpory a spolehlivosti, jsou označeny jako zajímavé a vypsány na výstup.

Tyto charakteristiky jsou počítány ze čtyřpolní tabulky 1. Číslo a uvádí, kolik záznamů splňuje předpoklad a zároveň i závěr. Číslo b je počet záznamů, které splňují předpoklad, ale nesplňují závěr. Číslo r vyjadřuje celkový počet záznamů, které splňují předpoklad. Číslo c , d a s vyjadřují analogicky počet záznamů, které ale nesplňují předpoklad. Číslo k (resp. l) vyjadřuje záznamy splňující (resp. nesplňující) závěr. Číslo n je celkový počet záznamů.

	Suc	$\neg$ Suc	Σ
Ant	a	b	r
$\neg$ Ant	c	d	s
Σ	k	l	n

Tabulka 1.: Frekvenční (čtyřpolní) tabulka.

Podpora je počet záznamů, které splňují předpoklad i závěr:

$$\text{supp}(Ant \Rightarrow Suc) = P(Ant \cup Suc) = \frac{a}{a + b + c + d}$$

Spolehlivost (někdy také platnost nebo správnost) je podmíněná pravděpodobnost závěru, pokud platí předpoklad:

$$\text{conf}(Ant \Rightarrow Suc) = P(Suc \mid Ant) = \frac{a}{a + b}$$

Existují také další charakteristiky, např. pokrytí (coverage) nebo zdvih (lift). Např. asociační pravidlo

$$Horcice \Rightarrow Parky [10 \%, 85 \%]$$

říká, že hořčice a párky byly koupeny dohromady v 10 % případů. Spolehlivost nám říká, že 85 % košíků, které obsahovaly hořčici, obsahovaly také párky.

2.2.1. Algoritmy hledání asociačních pravidel

Pro hledání asociačních pravidel, které splňují minimální hodnoty podpory a spolehlivosti, můžeme využít jednoduchý algoritmus generující všechny možné kombinace. V praxi se však hojně využívá *apriori algoritmus*.

Silné asociační pravidlo je takové pravidlo, které splňuje požadované minimální hodnoty podpory p_p i spolehlivosti s_p . Množina položek M se nazývá *frekventovaná*, jestliže dosahuje minimální hodnoty podpory, tzn. $P(M) \geq p_p$. Výsledkem hledání je seznam všech silných asociačních pravidel.

Jednoduchý algoritmus by mohl postupně generovat všechny možné kombinace predikátů na levé i pravé straně a testovat taková pravidla v datech. Takový algoritmus má však exponenciální časovou složitost a je tedy nepoužitelný, protože datová matice má většinou velký počet řádků i sloupců.

Apriori algoritmus

Tzv. *apriori algoritmus* využívá apriorní znalost silných asociačních pravidel [7]. Algoritmus probíhá ve dvou krocích. V první fázi jsou vyhledány frekventované množiny položek a ve druhé fázi jsou z nich generovány silná asociační pravidla.

Apriori znalost znamená, že pokud množina k položek není frekventovaná, potom žádná její nadmnožina také není frekventovaná. Přidáváním položek do M tedy pravděpodobnost $P(M)$ monotónně klesá. V algoritmu se využívá toho, že každá neprázdná podmnožina R frekventované množiny M musí být také frekventovaná.

Algoritmus funguje následovně:

- Vyhledá se frekventovaná množina položek (označíme ji L_1)

- Pomocí množiny L_1 se vyhledá L_2 , pomocí L_2 se vyhledá L_3 atd. dokud nelze nalézt žádnou další frekventovanou množinu L_k velikosti k . Hledání každé množiny L_k vyžaduje jeden průchod daty.

Před samotným uvedením algoritmu je nutné ještě zavést pojem *kandidát na frekventovanou množinu položek*. Je to množina položek, která nebyla prohlášena za nefrekventovanou, ale frekventovanost u ní nebyla potvrzena průchodem dat.

Nalezení L_k z L_{k-1} se provádí následovně:

- Z L_{k-1} vygenerujeme všechny kandidáty na L_k (označíme je C_k) a budeme předpokládat, že jsou položky v množinách lexikograficky seřazeny. Pak stačí kombinovat takové dvojice množin $l_i, l_j \in L_{k-1} (i \neq j)$, které mají $k-2$ prvních položek stejných a $(k-1)$ -položka je u l_i lexikograficky menší než u l_j . Výsledný kandidát vznikne spojením l_i s poslední $(k-1)$ -položkou l_j .
- Jelikož platí $L_k \subseteq C_k$, je nutné v C_k nalézt nefrekventované množiny a odstranit je. Nefrekventovanost množiny lze zjistit jedním průchodem dat. Množina C_k však může být rozsáhlá, a tak by bylo její zpracování pomalé. Zde můžeme opět využít apriorní znalost a vyloučit všechny množiny, pro které platí, že všechny jejich podmnožiny nejsou frekventované. Nakonec se s tímto redukováným C_k provede průchod daty.

Nalezení všech množin L_k je následující:

1. Vygeneruje se množina C_1 kandidátů c_i^1 na frekventované množiny velikosti jedna: $C_1 = \{c_i^1 : c_i^1 \subseteq I \wedge |c_i^1| = 1\}$, kde I je množina všech položek v datech. Nyní se provede průchod daty a pro každého kandidáta se zjišťuje počet výskytů. Vyloučením kandidátů, kteří nemají dostatečnou podporu p_p získáme množinu L_1 frekventovaných množin l_i^1 velikosti jedna: $L_1 = \{l_i^1 : l_i^1 \in C_1 \wedge P(l_i^1) \geq p_p\}$.
2. Nyní se vytvoří kandidáti délky dva vzájemným kombinováním frekventovaných množin délky jedna pomocí algoritmu uvedeného výše.
3. Tento postup opakujeme dokud nenastane, že $L_k = \emptyset$.

Ze všech nalezených frekventovaných položek nyní můžeme vygenerovat silná asociační pravidla.

1. Pro každou frekventovanou množinu položek l se vygenerují všechny neprázdné podmnožiny.
2. Pro každou neprázdnou podmnožinu s množiny l se otestuje nerovnost $\frac{\text{pocet_vyskytu}(l)}{\text{pocet_vyskytu}(s)} \geq s_p$ na splnění minimální spolehlivosti. Pokud je tato nerovnost splněna, patří toto asociační pravidlo $s \Rightarrow (l - s)$ mezi řešení.

2.3. Metoda GUHA

Metoda GUHA (General Unary Hypothesis Automaton - automat na obecné unární hypotézy) je metodou explorační analýzy českého původu. Metodu vytvořila skupina vědců kolem Petra Hájka. Její počátky sahají do 60. let 20. století. S podobnou myšlenkou přišla po třiceti letech i skupina kolem Agrawala a pojmenovali ji analýza nákupního košíku se svým algoritmem apriori viz kapitola 2.2.1. Tato metoda však nebyla tolik propracovaná jako metoda GUHA.

Cílem metody GUHA je poskytnout zajímavé vztahy (hypotézy) z analyzovaných (empirických) dat. Samotná metoda je realizována pomocí tzv. GUHA procedur. Na vstupu procedury jsou analyzovaná data a definice množiny potenciálně zajímavých vztahů. Výstupem procedury jsou všechny platné vztahy, které nejsou logicky odvoditelné z některého jiného výstupního vztahu.

Pojmy asociační pravidlo a hypotéza se volně zaměňují, protože se v podstatě jedná o to samé.

2.3.1. Pojmy

Obor	ChytryTelefon	Notebook	Pozornost
Informatika	Ne	Ne	vysoka
Apl. informatika	Ne	Ano	nizka
Informatika	Ano	Ano	vysoka
Informatika	Ne	Ano	nizka
Apl. informatika	Ne	Ne	vysoka

Tabulka 2.: Příklad tabulky dat studentů a jejich pozornosti.

Příklad pravidla vygenerovaného metodou GUHA z tabulky 2.:

$$Obor(Informatika) \wedge \neg ChytryTelefon(Ano) \Rightarrow_{0,9;0,7} Pozornost(Vysoka)$$

Syntaktické pojmy

Predikát - symbolické jméno veličiny (např. Obor, Pozornost)

Formule - jeden nebo více predikátů složených pomocí logických spojek negace, konjunkce, disjunkce (např. $Obor(Informatika) \wedge \neg ChytryTelefon(Ano)$)

Kvantifikátor - symbol určující druh a intenzitu souvislosti (např. $\Rightarrow_{0,9;0,7}$ - fundovaná implikace viz 2.3.2.)

Antecedent - formule na levé straně kvantifikátoru neboli předpoklad (např. $Obor(Informatika) \wedge \neg ChytryTelefon(Ano)$)

Sukcedent - formule na pravé straně kvantifikátoru neboli závěr (např. $Pozornost(Vysoka)$)

Formální sentence - zápis ve tvaru

$$formule1 \ q \ formule2$$

jehož pravdivost se v datech testuje (q je kvantifikátor) (např. $Obor(Informatika) \wedge \neg ChytryTelefon(Ano) \Rightarrow_{0,9;0,7} Pozornost(Vysoka)$)

Sémantické pojmy

Pravdivá sentence - sentence pravdivá v datech (funkce kvantifikátoru pro ni vrátila hodnotu 1)

Vstupem metody GUHA jsou data ve stejném formátu jako u asociačních pravidel (datová matice).

2.3.2. Kvantifikátory

Metoda GUHA systematicky generuje ze zadaných množin antecedentů a sukcedentů hypotézy, které testuje v datech. Na výstup se však zařazují pouze sentence, které podporují nějakou hypotézu. Kvantifikátory popisují druh a intenzitu hypotéz (tedy druh a intenzitu mezi antecedentem a sukcedentem) [8].

Druhy kvantifikátorů:

- Implikační - A (asi, většinou) je příčinou B ($A \Rightarrow B$)
- Asociační - A (asi, většinou) souvisí s B ($A \sim B$)
- Korelační - A a B (asi, většinou) spolu korelují za podmínky F ($A \text{ corr } B/F$)

GUHA používá jako své charakteristiky frekvence. Každý kvantifikátor je funkce frekvencí a, b, c, d čtyřpolní tabulky viz tab. 1. Pokud je tedy výsledkem zobrazení hodnota 1, patří sentence mezi řešení.

Výběr zajímavých kvantifikátorů:

- $\Rightarrow_{s,p}$ fundovaná implikace (pro $s \in (0; 1)$ a $p > 0$):

$$\Rightarrow_{s,p}(a, b, c, d) = 1, \text{ je-li } a \geq p \text{ a } a \geq s(a + b)$$

Tento kvantifikátor představuje požadavek na splnění minimální podpory s a spolehlivosti p v datech. Číslo p je zde však absolutní počet výskytů.

- $\Leftrightarrow_{s,p}$ dvojité fundovaná implikace (pro $s \in (0; 1)$ a $p > 0$):

$$\Leftrightarrow_{s,p} (a, b, c, d) = 1, \text{ je-li } a \geq p \text{ a } a \geq s(a + b + c)$$

- $\equiv_{s,p}$ fundovaná ekvivalence (pro $s \in (0; 1)$ a $p > 0$):

$$\equiv_{s,p} (a, b, c, d) = 1, \text{ je-li } a \geq p \text{ a } (a + d) \geq s(a + b + c + d)$$

2.3.3. Procedury metody GUHA

Metoda GUHA se skládá z několika procedur, mezi něž patří ASSOC, IMPL, COLLAPS a další.

Procedura ASSOC

Procedura ASSOC hledá v datech pravidla s asociačními kvantifikátory. Antecedent i sukcedent pravidel je ve tvaru elementární konjunkce.

Procedura IMPL

Procedura IMPL hledá pravidla s implikačními kvantifikátory. Antecedent je ve tvaru elementární konjunkce a sukcedent ve tvaru elementární disjunkce.

Procedura COLLAPS

Procedura COLLAPS hledá pravidla ve tvaru elementární konjunkce, pro které je podmíněná korelace dvou vybraných reálných veličin v datech vysoká.

2.4. Shluková analýza

Úkolem shlukové analýzy neboli shlukování je rozložit množinu objektů popsaných atributy do skupin, kde objekty ve stejné skupině jsou si navzájem více podobné než objekty z jiných skupin. Těmto skupinám říkáme shluky. Za shluk můžeme považovat množinu objektů ve skupině, které mají mezi sebou malé vzdálenosti. O shlucích není před začátkem shlukování nic známo. Shlukování má také za cíl výsledné shluky charakterizovat, případně určit reprezentanty shluků.

Shlukovou analýzu tvoří pouze jedna metoda, ale skládá se z několika různých metod. Tyto metody se mohou lišit v tom, co považují samy za shluk, svou složitostí a svými výstupy.

U shlukové analýzy vycházíme z toho, že dva objekty jsou si tím podobnější, čím více jejich atributů nabývá stejných nebo podobných hodnot. Každý objekt je vektor s n atributy. Podobnost dvou objektů si nejlépe představíme v geometrické interpretaci vektorů, kde objekty jsou si tím podobnější, čím jsou k sobě blíže.

Pro měření vzdálenosti dvou objektů a shluků je potřeba definovat metriku. Nechť M je množina. Metrika na M je reálná funkce $d : M \times M \rightarrow \mathbb{R}$, která každé uspořádané dvojici (x, y) bodů z M přiřazuje právě jedno reálné číslo $d(x, y)$ tak, že pro každé $x, y, z \in M$ platí:

1. $d(x, y) \geq 0$ (axiom nezápornosti)
2. $d(x, y) = 0 \Leftrightarrow x = y$ (axiom totožnosti)
3. $d(x, y) = d(y, x)$ (axiom symetrie)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (trojúhelníková nerovnost)

Množinu M spolu s metrikou d nazýváme metrickým prostorem, ozn. (M, d) . Prvky množiny M nazýváme body metrického prostoru (M, d) . Číslo $d(x, y)$ nazýváme vzdáleností bodu x, y .

Mezi nejznámější metriky patří Euklidovská a Manhattanská. Euklidovská metrika je definována na množině R^n (množina uspořádaných n -tic reálných čísel) předpisem:

Nechť $x, y \in R^n$, $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$, pak

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Existuje také čtvercová Euklidovská vzdálenost, která přiřazuje větší váhu bodům, které jsou si vzdálenější. Tato vzdálenost však není metrikou, protože nesplňuje trojúhelníkovou nerovnost. Její předpis je následující:

$$d^2(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

Manhattanská metrika byla inspirována pravoúhlou uliční sítí na Manhattanu. Tato metrika je definována na množině R^n předpisem:

Nechť $x, y \in R^n$, $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$, pak

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$

Volba metriky ovlivňuje výsledný tvar shluků a také některé objekty mohou být k sobě bližší nebo naopak vzdálenější.

Shlukování můžeme rozdělit dle cíle na hierarchické a nehierarchické.

- *Hierarchické shlukování* produkuje hierarchii shluků, kde každý následující rozklad je zjemněním předchozího rozkladu.
- *Nehierarchické shlukování* nevytváří hierarchii shluků. Tento typ rozkládá danou množinu na podmnožiny dle předem daného kritéria.

2.4.1. Hierarchické shlukování

Hierarchické shlukování je metoda shlukové analýzy, která vytváří hierarchii shluků [9]. Hierarchické shlukování se dělí na dva typy:

- *Divizivní*: přístup shora dolů. Na počátku tvoří všechny objekty jeden shluk. Shluky se postupně dělí a na konci tvoří každý shluk pouze jeden objekt.
- *Aglomerativní*: přístup zdola nahoru. Každý objekt tvoří na počátku jeden shluk. Shluky se postupně slučují a vytvářejí větší shluky. Nakonec se vytvoří jeden shluk obsahující všechny objekty.

Nevýhodou hierarchického shlukování je jeho časová složitost, která je nejméně kvadratická. Toto omezení je znát hlavně u velkých datasetů.

Výsledek hierarchického shlukování je většinou prezentován jako graf zvaný dendrogram. U aglomerativní metody hierarchického shlukování se vytváří velké množství hladin shluků. Konkrétně z n objektů se vytvoří $n - 1$ hladin shluků. Z tohoto velkého množství rozkladů je vhodné vyhledat pouze několik konkrétních. Existují dva základní způsoby určení výsledné hladiny:

- *hladina se zadaným počtem shluků* - pokud známe dopředu počet shluků, vybereme tuto hladinu.
- *největší rozdíly mezi sousedními hladinami* - takové hladiny určují nejlepší shluky, na grafu jsou znázorněny delšími vertikálními čarami.

Nepodobnost shluků

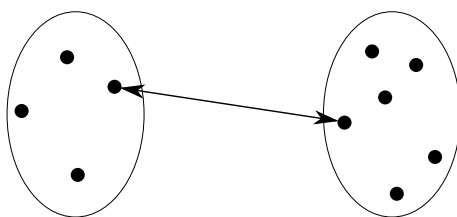
Pro určení dvou nejbližších shluků, které se mají sloučit (aglomerativní metoda) nebo rozdělit (divizivní metoda), se používá míra nepodobnosti [10]. Tato míra se získá kombinací metriky pro výpočet vzdálenosti shluků a shlukovací strategie.

Na počátku aglomerativní metody tvoří každý objekt jeden shluk. Vzdálenost mezi těmito shluky určuje metrika (např. Euklidovská). Jednotlivé vzdálenosti mezi shluky můžeme zaznamenat do čtvercové matice se všemi vzájemnými vzdálenostmi. Tato matice je symetrická a na diagonále jsou nuly.

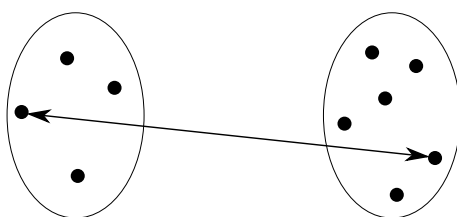
Shlukovací strategie se používá pro určení vzdáleností dvou shluků, které se skládají z více než jednoho objektu. Mezi nejvíce používané strategie patří tyto [11]:

- *Nejbližší soused (single linkage)*: vzdálenost dvou nejbližších objektů z různých shluků viz obr. 2.
- *Nejvzdálenější soused (complete linkage)*: vzdálenost dvou nejvzdálenějších objektů z různých shluků viz obr. 3.

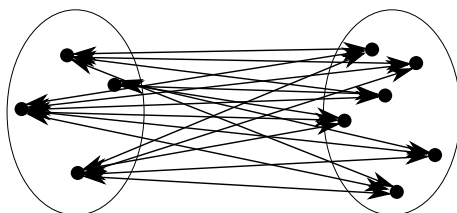
- *Průměrná vzdálenost objektů (average linkage)*: průměr vzdáleností všech dvojic objektů z různých shluků viz obr. 4.
- *Centroidní metoda (centroid linkage)*: vzdálenost center shluků, kde centroid je geometrický střed objektů shluku vypočítaný aritmetickým průměrem viz obr. 5.
- *Wardova metoda (Ward's linkage)*: suma čtverců vzdáleností center shluků.



Obrázek 2.: Metoda nejbližšího souseda



Obrázek 3.: Metoda nejvzdálenějšího souseda

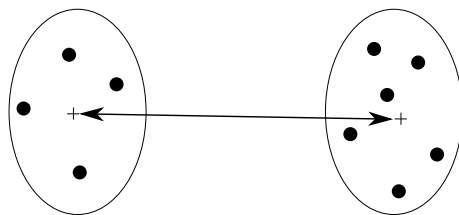


Obrázek 4.: Metoda průměrné vzdálenosti

Po sloučení nebo rozdělení dvou shluků není potřeba znovu počítat všechny vzdálenosti mezi jednotlivými shluky, ale pouze vzdálenosti změněných shluků od ostatních.

Jednoduchý algoritmus aglomerativní metody by vypadal následovně:

1. Vypočti matici vzdáleností mezi všemi objekty.
2. Nechť každý objekt je shluk.



Obrázek 5.: Centroidní metoda

3. Sluč dva nejbližší shluky.
4. Uprav matici vzdáleností.
5. Opakuj krok 3. pokud existuje více než jeden shluk.

2.4.2. Nehierarchické shlukování

Nehierarchické metody shlukování se snaží nalézt jediný optimální rozklad objektů na podmnožiny dle předem daného kritéria. V této skupině lze metody dále rozlišit dle jejich přístupu na:

- *Optimalizační k -středové*: iteračním způsobem zlepšují počáteční rozklad objektů.
- *Analýzy modů*: metody hledající nej hustší místa prostoru - mody.
- *Fuzzy k -středové*: pracují podobně jako iterační metody, ale pro každý objekt se navíc počítá, s jakou pravděpodobností patří do jakého shluku.
- *Neuronové sítě*: podobné objekty jsou reprezentovány blízkými si neurony v dané topologii sítě.

Nalezení optimálního rozkladu vyzkoušením všech možných kombinací shluků není díky množství objektů možné. Většina algoritmů tedy začíná s nějakým počátečním rozkladem, který se následně optimalizuje. Optimální počet shluků je nutné na začátku shlukování najít. Dle toho se metody dále dělí na metody s pevným počtem shluků nebo s proměnným počtem shluků, kde se počet mění v závislosti na řídicích parametrech.

Nejznámější nehierarchický algoritmus je k -means, který pracuje s centry shluků. Mezi další patří k -medoids, který pracuje s medoidy. Medoid neboli typický zástupce je objekt shluku, který má minimální průměrnou vzdálenost ke všem ostatním objektům ve shluku. Medoid je narození od centra shluku jedním z jeho reprezentantů. Algoritmy implementující tento přístup jsou např. PAM (Partitioning Around Medoids) nebo CLARA (Clustering LARge Applications).

k-means

Algoritmus k -means patří mezi optimalizační k -středové algoritmy s pevným počtem shluků. Objekty rozděluje do k shluků. Toto číslo k se zadává před shlukováním a musí být menší nebo rovno než počet objektů. Algoritmus poté náhodně vybere k počátečních centroidů reprezentujících jednotlivé shluky. Alternativně můžeme místo čísla k zadat k počátečních centroidů sami. k -means pracuje tak, že v každé iteraci přiřadí objekt do shluku, jehož geometrický střed vypočítaný dle aritmetického průměru je nejbližší. Algoritmus se tak snaží minimalizovat sumu čtverců uvnitř shluků. Po každém takovém přiřazení se znovu spočítá střed shluku. Nevýhoda algoritmu k -means je skutečnost, že počáteční centroidy shluků jsou vybrány náhodně nebo je musíme zadat. Při ručním výběru počátečních centroidů je vhodné vybrat objekty, které jsou od sebe dostatečně vzdálené, aby algoritmus vyprodukoval dobře rozlišitelné shluky.

Jednoduchý k -means algoritmus:

1. Vyber k objektů jako počáteční centroidy shluků.
2. Každý objekt přiřaď do shluku, jehož střed je nejbližší.
3. Vypočti centroidy shluků.
4. Opakuj krok 2. dokud se centroidy pohybují.

3. Software

3.1. LISp-Miner

LISp-Miner je akademický softwarový systém implementací několika data miningových procedur založených na metodě GUHA [12]. Systém běží pod operačním systémem Windows. LISp-Miner je vyvíjený skupinou pedagogů a studentů na Vysoké škole ekonomické v Praze. Hlavní GUHA procedurou je *4ft-Miner* pro různé typy asociačních pravidel. Mezi další procedury patří KL-Miner, CF-Miner, ETree-Miner atd. Následující text navazuje na kapitulu 2.3. o metodě GUHA.

3.1.1. Struktura systému

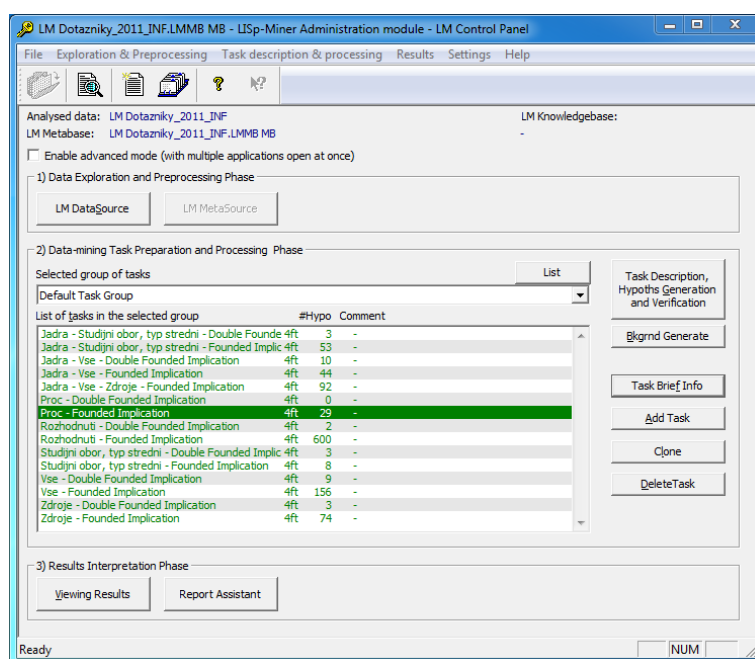
Systém LISp-Miner tvoří několik samostatných modulů pro přípravu dat, zadání úloh a zobrazení výsledků. Analyzovaná data jsou uložena v databázi MS Access. Zadání úloh a jejich výsledky jsou jako metadata uloženy také v databázi MS Access (metabáze).

Modul *LMAdmin.exe* se spouští jako první a kontroluje metabázi a analyzovaná data. Tímto modulem se také vytváří nová metabáze.

Modul *LMDataSource.exe* slouží pro přípravu a seznámení se s daty. V tomto modulu se zadávají atributy, ze kterých se budou dobývat asociační pravidla. Nezbytné je také zvolení primárního klíče, jehož hodnoty musí být unikátní. Modul umožňuje také vytvářet nové odvozené atributy. V neposlední řadě umožňuje analýzu dat pomocí čtyřpolních tabulek.

Modul *4ftTask.exe* je jedním z několika modulů, který implementuje metodu GUHA. Mezi další patří např. SD4ft-Miner, KL-Miner nebo CF-Miner. Práce s tímto modulem probíhá tak, že uživatel vytváří jednotlivé úlohy. V každé úloze lze nastavit všechny potřebné parametry. Po vygenerování se výsledky zobrazí v jednoduché textové podobě na výstupu a uloží se do metabáze.

Modul *4ftResult.exe* je jedním z několika modulů pro zobrazování výsledků dolování dat. Tento modul umožňuje zobrazovat výsledky pouze z modulu *4ftTask.exe*. Podobně pouze modul *SD4ftResult.exe* zobrazuje výsledky z modulu *SD4ftTask.exe*. Uživatel si může zobrazit výsledky v různých formách (grafy, čtyřpolní tabulky) a exportovat je do různých formátů (např. do HTML pomocí projektu SEWEBAR).



Obrázek 6.: LISp-Miner - Modul 4ftTask.exe

3.1.2. Procedura 4ft-Miner

Procedura *4ft-Miner* je implementace metody GUHA k získávání asociačních pravidel z dat [13]. Tato procedura je nejvíce používanou implementací procedury ASSOC.

Procedura *LMDataSource.exe* převede databázovou tabulku na analyzovanou matici dat. Procedura *4ft-Miner* pracuje s booleovskými atributy ze sloupců analyzované matice. Každý sloupec matice může obsahovat pouze konečné množství hodnot (nazýváme je kategorie).

Vstupem procedury *4ft-Miner* je:

- matice dat,
- množina asociačních pravidel, které mají být generovány,
- další parametry.

Relevantní otázky jsou asociační pravidla, která mají být generována a ověřena. Množina asociačních pravidel, která mají být generována a ověřena, se nazývá *množina relevantní otázek*. Ta je dána:

- množinou relevantních antecedentů,
- množinou relevantních sukcedentů,
- množinou relevantních podmínek,
- 4ft-kvantifikátorem.

Antecedent, sukcedent a podmínka se skládá z konjunkcí *dílčích cedentů*, což jsou konjunkce nebo disjunkce literálů. *Literál* je booleovský atribut $A(\alpha)$ nebo jeho negace $\neg A(\alpha)$ základního booleovského atributu. Množina α v literálu se nazývá koeficient. Tato množina obsahuje jen přípustné hodnoty atributu A . Literál $A(\alpha)$ se nazývá *pozitivní literál*, $\neg A(\alpha)$ se nazývá *negativní literál*. *Délka literálu* je definována jako počet kategorií v koeficientu α .

Konjunkce dílčích antecedentů se nazývá *relevantní antecedent*. Např.

$$\varphi = \varphi_1 \wedge \varphi_2 \wedge \dots \wedge \varphi_k$$

Množina relevantních antecedentů je zadána alespoň jednou množinou relevantních dílčích antecedentů a zadáním minimální a maximální délky antecedentu. Součet délek dílčích antecedentů tvoří celkovou délku antecedentu. Délka dílčího antecedentu je dána počtem literálů v tomto antecedentu. Délka dílčího antecedentu může být i 0, což vede na konjunkce a disjunkce délky 0, které jsou vždy pravdivé. Množina relevantních antecedentů je zadána především:

- minimální a maximální délkou dílčího antecedentu,
- seznamem atributů, z nichž jsou generovány literály,
- použitou booleovskou spojkou (konjunkce, disjunkce),

- definicí množiny literálů, které mají být pro každý atribut automaticky generovány.

Množina literálů, které mají být pro každý atribut generovány, je dána:

- typem koeficientu (např. *subsets*, *intervals*, *cyclic intervals*, *left cuts*, *right cuts*, *cuts*, *one particular category*),
- minimální a maximální délkou literálu,
- následujícím typem:
 - generovat pouze pozitivní literály,
 - generovat pouze negativní literály,
 - generovat pozitivní i negativní literály.

Zadání množiny relevantních sukcedentů a relevantních podmínek je totožné jako v případě relevantních antecedentů. Zadání 4ft-kvantifikátoru probíhá výběrem ze seznamu kvantifikátorů (některé byly uvedeny v kapitole 2.3.2.).

Výstupem procedury *4ft-Miner* jsou všechna prostá asociační pravidla. *Prosté asociační pravidlo* je pravidlo, které je pravdivé a zřejmým způsobem neplyne z jiného pravidla uvedeného na výstupu.

3.2. R

R je programovací jazyk, otevřený a multiplatformní software určený především pro statistickou analýzu a její grafické zpracování [14]. R je implementací programovacího jazyka S. Kód napsaný v jazyce S lze většinou bez problémů spustit přímo v R. Práce s R probíhá pomocí příkazové řádky, ale jsou dostupná různá grafická rozhraní pro práci s ním.

Komunita kolem jazyka R se v posledních letech značně rozšířila a je aktivní. R podporuje procedurální programování, ale také objektově orientované programování. Části systému jsou napsány pomocí samotného jazyka R. Uživatel může při psaní kódu využívat také programovací jazyky C, C++ a Fortran.

R je lehce rozšiřitelné pomocí funkcí a rozšíření. Balíky funkcí rozšiřující R se nazývají *packages*. Velké množství balíků je již obsaženo v základní instalaci programu R. Další balíky jsou dostupné v různých repozitářích. Nejznámější repozitář balíků R je centrální repozitář *Comprehensive R Archive Network* (CRAN) na adrese <http://cran.r-project.org/>, který obsahuje několik tisíc balíků.

3.2.1. Asociační pravidla

Získávání asociačních pravidel v programu R zajišťuje balík *arules*. Tento balík obsahuje funkci *apriori*, která využívá stejnojmenný algoritmus. Použití funkce je následující:

```
apriori(data, parameter = NULL, appearance = NULL, control = NULL)
```

První parametr *data* slouží k zadání datové matice. Druhý parametr slouží k předání parametrů jako je podpora nebo spolehlivost. Parametrem *appearance* lze nastavit zobrazování pouze určitých pravidel. Posledním parametrem lze ovlivnit výkonnost algoritmu.

3.2.2. Shlukování

Při hierarchickém shlukování jsem využil balík *stats*. Samotné hierarchické shlukování pomocí aglomerativního přístupu provádí funkce *hclust*:

```
hclust(d, method = "complete", members=NULL)
```

První parametr *d* je matice vzdáleností. V druhém parametru se může nepovinně zvolit metoda shlukování (single, complete, average, centroid, ward). Při použití Wardovy metody je nutné předat čtvercovou Euklidovskou vzdálenost. Poslední parametr se využívá při shlukování z jiného než počátečního stavu, tzn. shlukování může začít třeba uprostřed dendrogramu.

Matice vzdáleností se získá pomocí funkce *dist*, která podporuje i chybějící hodnoty a proporčně je započítává do vzdálenosti.

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
```

První parametr *x* je číselná matice. Druhým parametrem lze zvolit metodu výpočtu vzdáleností mezi dvěma objekty. Parametr *diag* udává, zda se jedná o diagonální matici. Parametr *upper* udává, zda se mají na výstupu objevit i hodnoty v horním trojúhelníku matice. Posledním parametrem lze zvolit mocninu Minkowského vzdálenosti.

Ořezání výsledného stromu provádí funkce *cutree*, která požaduje objekt z funkce *hclust* a počet výsledných shluků (parametr *k*) nebo výšku ve stromě (parametr *h*), kde má být ořezán. Funkce vrací číselný vektor přiřazující každému objektu číslo shluku.

```
cutree(tree, k = NULL, h = NULL)
```

Pro zpracování výsledků hierarchického shlukování jsem napsal funkci *writeHierchClustersSummary*, která jako své parametry požaduje původní data, počet shluků, vektor přiřazující každému objektu číslo shluku a název výstupního souboru. Tato funkce zapíše jako CSV soubor (hodnoty jsou oddělené čárkou) tabulku obsahující v prvním sloupci jména jednotlivých sloupců původních dat. Dále tabulka obsahuje pro každý shluk jeden sloupec, ve kterém je uvedeno číslo shluku, počet objektů ve shluku a frekvence výskytů všech hodnot, které mohl sloupec v původních datech nabývat.

```
writeHierchClustersSummary(data, count, clusterNr, filename)
```

Pro spočítání procentuálního rozdílu objektů padnoucích do jiného shluku jsem napsal funkci *countCutsDiffInPercent*. První tři parametry funkce jsou objekty z funkce *hclust*. Parametr *start* určuje počáteční počet shluků a parametr *end* koncový počet shluků. Funkce postupně iteruje od hodnoty parametru *start* po *end*. Uvnitř cyklu se pomocí funkce *countVectorsDiff* spočítají rozdílné hodnoty ve vektorech vrácené funkcí *cutree*.

```
countCutsDiffInPercent(hierch1, hierch2, hierch3, start, end)
```

```
countVectorsDiff(vec1, vec2)
```

3.2.3. k-means

Shlukování pomocí algoritmu *k-means* provádí v R stejnojmenná funkce [15]. Prvním parametrem je datová matice. Za parametr *centers* se může dosadit počet shluků (číslo *k*) nebo již vybraná množina centroidů shluků. V případě dosazení čísla se centroidy shluků vyberou náhodně. Parametr *iter.max* udává maximální počet iterací algoritmu. Pokud je parametr *centers* číslo, můžeme parametrem *nstart* ovlivnit počet náhodně vybraných centroidů shluků. Poslední parametr *algorithm* umožňuje zvolit typ algoritmus.

```
kmeans(x, centers, iter.max = 10, nstart = 1, algorithm =  
c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))
```

4. Data mining dotazníků při zápisu do 1. ročníku PřF UP

Má diplomová práce se zabývá zjištěním zajímavých souvislostí na dotaznících vyplněných studenty při zápisu do 1. ročníku Přírodovědecké fakulty Univerzity Palackého v Olomouci v roce 2010 a 2011. Dotazníky z roku 2012 bohužel nejsou k dispozici. Cílem je analyzovat data oborů Aplikovaná Informatika a Informatika a dále data všech oborů a najít v nich zajímavé souvislosti.

4.1. Struktura dotazníků

Data z dotazníků byla uložena ve formátu Excel se sloupci studijní obor, střední škola, bydliště, kraj, počet přihlášek a dále 34 otázek. Tyto otázky byly rozděleny do tří tematických okruhů.

První okruh otázek se zabývá zjištěním, co ovlivnilo rozhodnutí studenta studovat na PřF UP v Olomouci. Otázky byly následující: rodiče; učitel/é na SŠ nebo ZŠ; osobní kontakt se studentem nebo zaměstnancem PřF UP; spolupráce na projektech PřF UP během studia na SŠ, účast na odborných soustředěních apod.; propagační materiály PřF UP (webové stránky, brožury apod.); exkurze na pracovišti PřF UP, účast na propagační akci pořádané PřF UP; možnost odstěhovat se od rodičů a být nezávislý/á; možnost zůstat v místě bydliště; chtěl/a jsem být se svými kamarády (partnerem); snadnost přijetí na PřF UP (např. bez přijímacích zkoušek); možnost podílet se na špičkovém výzkumu; relativně nízké životní náklady v Olomouci; prestiž a výsledky PřF UP; vize nenáročného studia na PřF UP; atraktivní studijní obor; fakt, že na jinou VŠ, kam jsem se hlásil/a, mě nepřijali a jiné důvody.

Druhý okruh otázek zjišťuje, odkud a z jakých informačních zdrojů se student dozvěděl o možnosti studia na PřF UP v Olomouci. Otázky byly následující: od spolužáků, známých, příbuzných apod.; od učitele/ů na střední škole; z televize, z rozhlasu, z novin; z internetových stránek PřF UP; na dni otevřených dveří nebo na jiných propagačních akcích pořádaných PřF UP; na veletrhu Gaudeamus; z propagační brožurky/letáku vydaného PřF UP; z Facebooku či jiné sociální sítě; z prezentace PřF UP na vaší SŠ; na odborných soutěžích, soustředěních apod.; z jiných informačních zdrojů.

Třetí okruh otázek obsahuje informace, proč se student rozhodl studovat daný obor. Otázky byly následující: abych mohl/a po absolvování školy získat lukrativní zaměstnání; chtěl/a bych se dát na vědeckou dráhu a dělat výzkum v této oblasti; je to snadná cesta, jak získat vysokoškolský diplom; obor mě zajímá a baví, chci se o něm dozvědět víc; chtěl/a bych dělat práci, pro kterou je potřeba tento obor vystudovat a jiné důvody.

Student mohl odpovědět na každou otázku odpovědí:

- určitě ano,
- spíše ano,
- spíše ne,
- určitě ne.

Každý sloupec otázky byl dále rozdělen do čtyř podsloupců s názvy určitě ano, spíše ano, spíše ne a určitě ne. Odpověď studenta byla do příslušného podsloupce zanesena hodnotou „x“.

U každé otázky a i celého dotazníku byla také možnost poznámky. Řádky tvořily jednotlivé odpovědi studentů. Poznámka z celého dotazníku byla uvedena v posledním sloupci.

Data z roku 2010 tvořilo celkem 577 záznamů z různých oborů. Obory Aplikovaná informatika a Informatika byly zastoupeny v počtu 79 záznamů. Data z roku 2011 tvořilo celkem 1124 záznamů z různých oborů, z toho 147 záznamů bylo z oboru Aplikovaná informatika a Informatika. U určitého počtu záznamů nebyly vyplněny odpovědi na všechny otázky.

4.2. Příprava dat

V metodice CRISP-DM viz kap. 2.1. následuje po porozumění datům fáze přípravy dat. Tato nejnáročnější fáze metodiky má za úkol připravit data pro zpracování samotnými metodami dolování dat.

U obou dat dotazníků jsem ručně prošel všechny záznamy oborů, které v mnohých případech obsahovaly jen číslo programu. Množství záznamů obsahuje pouze název programu bez specifikace oboru. Mezi daty z dotazníků roku 2010 a 2011 byly také různě pojmenovány stejné obory např. s různými mezerami. Tyto odlišnosti jsem v datech opravil a pomocí seskupení oborů ověřil, že data neobsahují žádné překlepy. V datech z let 2010 a 2011 se nachází celkem 85 programů a oborů.

Do všech podsloupců otázek jsem místo hodnoty „x“ doplnil hodnotu *UrciteAno*, *SpiseAno*, *SpiseNe* a *UrciteNe*. Následně jsem tyto podsloupce sloučil do jediného a pojmenoval jej podle příslušné otázky.

U záznamů z oborů Aplikovaná informatika a Informatika jsem do dat přidal sloupec *TypStredni*. Do tohoto sloupce jsem doplnil hodnoty *ss* (střední škola), *sps* (střední průmyslová škola) nebo *gymnazium* dle typu střední školy jednotlivých záznamů.

4.3. Rok 2011 - Apl. informatika a Informatika

Data z oborů Aplikovaná informatika a Informatika z dotazníků roku 2011 obsahují celkem 147 záznamů.

4.3.1. Asociační pravidla

Jako datová matice do programu R byl použitý CSV soubor, který jsem získal exportováním dat z formátu Excel. Sloupce datové matice byly následující: studijní obor, počet přihlášek, střední škola, typ střední školy a dále všechny otázky ze třech okruhů dotazníku. Hodnota sloupce otázky z okruhů dotazníku mohla obsahovat jednu z těchto hodnot: *UrciteAno*, *SpiseAno*, *SpiseNe*, *UrciteNe* případně prázdnou hodnotu.

Následně byly pomocí funkce *apriori* získány asociační pravidla se zadanou minimální podporou (*supp*) a spolehlivostí (*conf*). Při hledání asociačních pra-

videl jsem také pomocí parametru omezil maximální délku položek v pravidle na tři. Toto omezení se mi jevilo jako optimální, jelikož při větších hodnotách se generovala pravidla, která byla pouze kombinací již známých pravidel.

Příklad generování asociačních pravidel s nastavenými parametry:

```
rules <- apriori(dotazniky, parameter = list(supp=0.1, conf=0.5,
maxlen=3))
```

Pravidlo	Podpora	Spolehl.
{ } => {RozhodnutiAtraktivniObor=SpiseAno}	0,51	0,51
{ } => {RozhodnutiPrestizPrf=SpiseAno}	0,52	0,52
{ } => {RozhodnutiVyzkum=SpiseAno}	0,52	0,52
{ } => {ZdrojeTelevizeNoviny=UrcNe}	0,52	0,52
{ } => {ZdrojeGaudeamus=UrcNe}	0,52	0,52
{ } => {RozhodnutiUcitele=UrcNe}	0,56	0,56
{ } => {ProcZajemObor=UrcAno}	0,56	0,56
{ } => {ZdrojeFacebook=UrcNe}	0,56	0,56
{ } => {ZdrojePrezentaceNaSs=UrcNe}	0,58	0,58
{ } => {StudijniObor=Apl, informatika}	0,62	0,62
{ } => {RozhodnutiNeprijetiVs=UrcNe}"	0,70	0,7
{ } => {RozhodnutiSpolupraceProjekty=UrcNe}	0,72	0,72
{ } => {ZdrojeSoutezeSoustredeni=UrcNe}	0,71	0,71
{StredniSkola=VOŠ a SPŠE Ol.} => {TypStredni=sps}	0,11	1
{ProcDrahaVaV=UrcNe} => {RozhodUcitele=UrcNe}	0,10	0,79
{ZdrSpolZnami=UrcNe} => {RozhRodice=UrcNe}	0,11	0,73
{ZdrSpolZnami=UrcNe} => {ZdrUciteleSs=UrcNe}	0,12	0,82
{ZdrSpolZnami=UrcNe} => {ProcSnadnyDiplom=UrcNe}	0,10	0,68
{ZdrSpolZnami=UrcNe} => {RozhOsobKontakt=UrcNe}	0,13	0,86

Tabulka 3.: Vybraná asoc. pravidla (min. podpora 0,1; min. spolehlivost 0,5).

I přes nespočet pokusů s nastavením parametrů funkce *apriori* jsem získával velké množství asociačních pravidel, která byla již zřejmá nebo byla jejich kombinací viz 3. Vhodnějším nástrojem pro získání asociačních pravidel je metoda GUHA viz kapitola 4.3.4.

4.3.2. Hierarchické shlukování

Pro shlukování jsem opět využil program R. Při přípravě dat jsem převedl data z formátu Excel do formátu CSV s hodnotami oddělenými středníkem. Data obsahovala tyto sloupce: střední škola, typ střední školy, počet přihlášek, studijní

obor a dále 34 otázek ze tří tematických okruhů. Samotné shlukování jsem prováděl pouze na samostatných otázkách. Sloupce střední škola, typ střední školy, počet přihlášek a studijní obor jsem po získání výsledků shlukování pouze přiřadil výsledným objektům zpět. Tento formát dat je používán i pro následující úlohy hierarchického shlukování.

Před shlukováním jsem jednotlivým odpovědím přiřadil číselné hodnoty. Odpovědi určitě ano jsem přiřadil váhu 0; odpovědi spíše ano váhu 0,2; odpovědi spíše ne váhu 0,8 a odpovědi určitě ne váhu 1. Tyto číselné hodnoty jsem jednoduše nahradil za řetězcové ekvivalenty `UrciteAno`, `SpiseAno`, `SpiseNe` a `UrciteNe`. Při takto zvolených hodnotách mají odpovědi určitě ano a spíše ano k sobě blíže než k ostatním dvěma odpovědím. Pro nezodpovězené otázky jsou v datech prázdná místa, která se poté v programu R interpretují jako chybějící hodnoty.

Metriku jsem zvolil euklidovskou. Při použití manhattanské metriky byly výsledky velmi podobné.

Na data jsem postupně aplikoval čtyři metody shlukování. Metoda `single linkage`, jak je na dendrogramu 7. patrné, vytvořila strom, který téměř v každé hladině obsahuje dva shluky. První shluk tvoří jediný objekt a druhý shluk tvoří zbývající objekty. Metoda `average linkage` vytvořila velice podobný dendrogram 8., který už však nemá v téměř každé hladině shluk obsahující jediný objekt, ale více objektů. V dalších kapitolách se u těchto dvou metod již neodkazuje na jejich dendrogramy z důvodu velmi podobných výsledků. Metoda `complete linkage` vytvořila dendrogram 9., který je na první pohled dobře strukturovaný. Nakonec Wardova metoda vytvořila dendrogram velice podobný předchozí metodě viz 10., kde jsou však dobře patrné jednotlivé shluky. Pro bližší zkoumání jsem si vybral metodu `complete linkage` a Wardovu metodu, jelikož ostatní dvě metody nevytvořily příliš vhodné dendrogramy.

U metody `complete linkage` jsem zvolil hladinu se čtyřmi shluky. Typičtí zástupci shluků (medoidi) jsou uvedeni v tabulce 44.

První shluk obsahuje 56 objektů, druhý 19 objektů, třetí 66 objektů a čtvrtý shluk 6 objektů. První shluk obsahuje část studentů, kteří se rozhodovali podle osobního kontaktu se studentem nebo zaměstnancem PřF UP v Olomouci (24 objektů s odpověďmi určitě ano a spíše ano).

Druhý shluk obsahuje studenty, kteří se nerozhodovali podle osobního kontaktu. Více než polovina studentů se nerozhodovala podle propagačních materiálů PřF UP (13 objektů s odpověďmi určitě ne a spíše ne), naopak se rozhodovali podle možnosti zůstat v místě svého bydliště (11 objektů), snadnosti přijetí na PřF UP (15 objektů) a nízkých životních nákladů v Olomouci (12 objektů).

Třetí shluk obsahuje studenty, kteří se stejně jako u prvního shluku rozhodovali podle osobního kontaktu (25 objektů). Velká část studentů se nerozhodovala podle propagačních materiálů PřF UP (24 objektů).

Čtvrtý shluk obsahuje studenty, kteří se pouze rozhodovali podle osobního kontaktu, dle možnosti odstěhování se od rodičů, snadnosti přijetí na PřF UP, možnosti podílet se na výzkumu, nízkých životních nákladů v Olomouci a vize ne-

náročného studia. Většina studentů v tomto shluku se nerozhodovala podle možnosti zůstat v místě svého bydliště (4 objekty). Studenti se dozvěděli o možnosti studia od spolužáků a známých, z internetových stránek PřF UP, dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Studenti se rozhodli studovat kvůli vědecké dráze a možnosti dělat výzkum, a také kvůli snadnému získání vysokoškolského diplomu.

U Wardovy metody jsem použil hladinu se třemi shluky. Každý shluk obsahuje podobné množství objektů. V prvním shluku se nachází 48 objektů, ve druhém 54 objektů a ve třetím 45 objektů. Typičtí zástupci shluků jsou uvedeny v tabulce 45.

První shluk obsahuje část studentů, kteří se rozhodovali podle možnosti odstěhování od rodičů a naopak se nerozhodovali podle možnosti zůstat v místě bydliště.

Druhý shluk obsahuje studenty, kteří se víceméně nerozhodovali podle osobního kontaktu se studentem nebo zaměstnancem PřF UP a podle prestiže PřF UP. Velká část studentů se rozhodovala dle snadnosti přijetí na PřF UP a také se dozvěděla o studiu od spolužáků a známých.

U třetího shluku se vyskytuje více studentů, kteří se rozhodovali podle rodičů, učitelů na SŠ nebo ZŠ a propagačních materiálů PřF UP. Značná část studentů se také rozhodovala podle exkurze na pracovišti PřF UP, o studiu se dozvěděli od učitelů na SŠ, na dni otevřených dveří a z propagační brožury. Část studentů tohoto shluku se také rozhodla studovat pro snadné získání VŠ diplomu.

Váhy odpovědí

Při volbě vah odpovědí vzniká otázka, zda budou objekty tvořit stejné shluky i při odlišných hodnotách vah. Pro ověření výsledků hierarchického shlukování jsem přiřadil odpovědím spíše ano a spíše ne mírně odlišné váhy.

V prvním případě jsem přiřadil odpovědi spíše ano váhu 0,25 a odpovědi spíše ne váhu 0,75. Ve druhém případě jsem přiřadil odpovědi spíše ano váhu 0,3 a odpovědi spíše ne váhu 0,7.

Změna těchto vah se projeví v případech, kdy odpovědi studentů obsahují více hodnot spíše ano a spíše ne. V takových situacích si pak jsou více nebo méně vzdáleni v závislosti na počtu hodnot těchto odpovědí.

Pro demonstrační účely jsem vytvořil jednoduché dotazníky se čtyřmi otázkami, dvěma studenty a jejich různými odpověďmi. Poté jsem vypočítal Euklidovskou a Manhattanskou vzdálenost těchto dvou objektů. V tabulce 4. je uveden jednoduchý případ s jednou opačnou odpovědí. V tabulce 5., 6. a 7. jsou uvedeny případy se třemi různými odpověďmi a různými vahami pro odpovědi. V tabulce 5. je vidět, že Euklidovská vzdálenost je větší než v případě jedné opačné odpovědi uvedené v tabulce 4. V tabulkách 6. a 7. jsou Euklidovské vzdálenosti menší než v tabulce 4. Na těchto dvou případech je viditelné, že malá změna vah odpovědí vede ke změně vzdáleností mezi objekty s různými odpověďmi.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem

	Otázka č. 1	Otázka č. 2	Otázka č. 3	Otázka č. 4
Student 1	0	0	0	1
Student 2	0	0	0	0
Euklidovská vzd.:	1			
Manhattanská vzd.:	1			

Tabulka 4.: Tabulka odpovědí a vzdáleností s jednou opačnou odpovědí.

	Otázka č. 1	Otázka č. 2	Otázka č. 3	Otázka č. 4
Student 1	0,2	0,2	0,2	0
Student 2	0,8	0,8	0,8	0
Euklidovská vzd.:	1,04			
Manhattanská vzd.:	1,8			

Tabulka 5.: Tabulka odpovědí a vzdáleností se třemi různými odpověďmi.

opět aplikoval hierarchické shlukování se čtyřmi různými metodami. V tabulce 8. a 9. jsou uvedeny procentuální podíly jednotlivých vah metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Téměř všechny podíly metody complete linkage činily více než 50 %. U vah #1 a #2 jsou v případě dvou shluků podíly pod 1 %. Ve Wardově metodě jsou podíly mezi vahami #1 a #2 menší než 50 %. Celkově pak jsou u této metody všechny procentuální podíly menší než u metody complete linkage.

Z toho je vidět, že určitý počet objektů se s různými vahami odpovědí nachází v jiném než původním shluku a ostatní objekty se nacházejí ve stejném shluku. Provedl jsem experiment, ve kterém jsem vzal objekty, které se nacházely ve stejném shluku mezi vahami 0; 0,2; 0,8 a 1 a 0; 0,25; 0,75 a 1 u Wardovy metody při hladině třech shluků. O těchto objektech můžeme mluvit jako o tzv. jádrech shluků. Naopak objekty, které se nacházejí v jiných shlucích, můžeme označit za tzv. fluktuanty. Na objektech jader jsem provedl opět hierarchické shlukování se stejnými metodami. U metody complete linkage lze v tabulce 10. spatřit menší procentuální podíl mezi vahami #1 a #2 u třech až šesti shluků. U Wardovy metody je v tabulce 11. rovněž vidět menší procentuální podíl mezi vahami #1 a #2.

V případě metody complete linkage jsem zvolil hladinu se dvěma shluky (dendrogram 11.). Typičtí zástupci shluků jsou uvedeny v tabulce 46.

	Otázka č. 1	Otázka č. 2	Otázka č. 3	Otázka č. 4
Student 1	0,25	0,25	0,25	0
Student 2	0,75	0,75	0,75	0
Euklidovská vzd.:	0,87			
Manhattanská vzd.:	1,5			

Tabulka 6.: Tabulka odpovědí a vzdáleností se třemi různými odpověďmi.

	Otázka č. 1	Otázka č. 2	Otázka č. 3	Otázka č. 4
Student 1	0,3	0,3	0,3	0
Student 2	0,7	0,7	0,7	0
Euklidovská vzd.:	0,69			
Manhattanská vzd.:	1,2			

Tabulka 7.: Tabulka odpovědí a vzdáleností se třemi různými odpověďmi.

První shluk obsahuje několik studentů, kteří se rozhodli studovat na PřF UP podle učitelů na SŠ nebo ZŠ, osobního kontaktu se studentem nebo zaměstnancem, propagačních materiálů PřF UP, exkurze na pracovišti, možnosti podílet se na výzkumu a prestiže PřF UP. Studenti z tohoto shluku se dozvěděli o možnosti studovat na PřF UP převážně od svých spolužáků a známých. Část studentů se rozhodla studovat kvůli snadnému získání VŠ diplomu.

Druhý shluk obsahuje studenty, kteří se nerozhodovali dle rodičů ani podle vize nenáročného studia.

U Wardovy metody jsem opět zvolil hladinu se dvěma shluky (dendrogram 12.). Typičtí zástupci shluků jsou uvedeny v tabulce 46.

První shluk obsahuje studenty, kteří se rozhodli studovat PřF UP kvůli odstěhování od rodičů.

Druhý shluk obsahuje část studentů, které ovlivnily exkurze na pracovišti PřF UP, možnost zůstat v místě svého bydliště a nízké životní náklady v Olomouci. O studiu se dozvěděli především od spolužáků a známých, učitelů na SŠ a ze dne otevřených dveří. Část studentů by se také chtěla dát na vědeckou dráhu a provádět výzkum.

Dále jsem provedl shlukování dle jednotlivých kategorií otázek v dotazníku. Ten obsahuje tři kategorie otázek. První kategorii „Co ovlivnilo Vaše rozhodnutí studovat na PřF UP v Olomouci?“ tvoří 17 otázek, druhou kategorii „Odkud (z jakých informačních zdrojů) jste se dozvěděl/a o možnosti studia na PřF UP.“ tvoří 11 otázek a třetí kategorii „Proč jste se rozhodl/a studovat daný obor?“

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	0,68 %	56 %	55 %
3	54 %	65 %	35 %
4	60 %	88 %	73 %
5	63 %	84 %	78 %
6	54 %	84 %	76 %

Tabulka 8.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	14 %	21 %	14 %
3	18 %	74 %	73 %
4	20 %	69 %	68 %
5	29 %	71 %	66 %
6	49 %	75 %	66 %

Tabulka 9.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

tvoří 6 otázek.

Na všechny tři kategorie otázek jsem aplikoval hierarchické shlukování se čtyřmi různými metodami. V tabulce 12. jsou vidět vysoké procentuální rozdíly mezi jednotlivými váhami u první kategorie otázek. U druhé kategorie otázek jsou procentuální rozdíly menší, jak je vidět v tabulce 13. Nejmenší procentuální rozdíly jsou u třetí kategorie otázek, viz tabulka 14.. Podobné procentuální rozdíly jsou také u Wardovy metody, viz tabulky 15., 16. a 17.

4.3.3. k-means

Algoritmus *k*-means jsem zkoušel několikrát na data aplikovat s různými parametry, ale jeho použití se nejevilo jako vhodné, jelikož musí být předem známe počáteční centroidy. Pokud se do algoritmu zadá jen počet výsledných shluků, algoritmus vybírá při každém spuštění náhodné centroidy. To má za následek jiné výsledky po každém spuštění algoritmu.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	23 %	71 %	73 %
3	25 %	69 %	69 %
4	27 %	68 %	71 %
5	58 %	70 %	71 %
6	49 %	57 %	63 %

Tabulka 10.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	6 %	8 %	12 %
3	6 %	68 %	68 %
4	6 %	68 %	68 %
5	7 %	64 %	63 %
6	7 %	77 %	76 %

Tabulka 11.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

4.3.4. Metoda GUHA

Data pro LISp-Miner tvoří databáze MS-Access, do které jsem importoval data z formátu Excel. Databáze obsahuje tabulky pro data z let 2010, 2011, sloučená data pro roky 2010 a 2011 a dále záznamy s obory Aplikovaná informatika a Informatika.

Pro generování pravidel jsem využil proceduru *4ft-Miner*. V první fázi jsem vytvořil metabázi pomocí modulu *LMDataSource.exe* a jako atributy, ze kterých se budou pravidla dobývat, jsem zvolil všechny dostupné. V modulu *4ftTask.exe* jsem následně zadával všechny úlohy.

V první úloze jsem do antecedentu zvolil všechny atributy ze tří okruhů otázek kromě těchto atributů: ProcJine, RozhodnutíJine a ZdrojeJine. Tyto atributy byly vynechány, jelikož dotazníky obsahovaly u těchto otázek velké množství odpovědí určitě ne. Tyto odpovědi pak generovaly mnoho pravidel, která jsem nepovažoval za zajímavá. Délku antecedentu jsem nastavil na nula až jedna. Tožným způsobem jsem nastavil i sukcedent, jehož délka mohla být jedna až dva. Jako kvantifikátor jsem nastavil BASE s hodnotou 5 % (nejméně 5 % záznamů musí splňovat předpoklad i závěr) a fundovanou implikaci s hodnotou 0,7. Takto zadaná úloha vygenerovala celkem 156 pravidel.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	69 %	79 %	22 %
3	66 %	76 %	27 %
4	87 %	81 %	51 %
5	54 %	81 %	77 %
6	56 %	91 %	77 %

Tabulka 12.: Tabulka proc. podílů shluků jednotlivých vah u complete linkage první kategorie.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	33 %	27 %	11 %
3	47 %	27 %	24 %
4	44 %	35 %	54 %
5	40 %	47 %	42 %
6	44 %	54 %	50 %

Tabulka 13.: Tabulka proc. podílů shluků jednotlivých vah u complete linkage druhé kategorie.

Ve výsledcích je antecedent pravidla uveden na levé straně značky $>:<$, sukcedent se nachází na pravé straně této značky. Výběr zajímavých pravidel:

- č. 1: ProcSnadnyDiplom(SpiseAno) $>:<$ ZdrojeSpoluzaciZnami(SpiseAno)
- č. 2: RozhodnutiExkurze(UrciteAno) $>:<$ RozhodnutiVyzkum(SpiseAno)
- č. 3: RozhodnutiPropagacniMaterialy(UrciteAno) $>:<$ ZdrojeInternetPrf(UrciteAno)
- č. 4: RozhodnutiRodice(SpiseAno) $>:<$ RozhodnutiVyzkum(SpiseAno)
- č. 5: ZdrojeDenOtevrenychDveri(UrciteAno) $>:<$ ProcZajemObor(UrciteAno)
- č. 6: ZdrojeFacebook(SpiseAno) $>:<$ RozhodnutiPrestizPrf(SpiseAno)
- č. 7: ZdrojeGaudeamus(SpiseAno) $>:<$ RozhodnutiPrestizPrf(SpiseAno)
- č. 8: ZdrojeGaudeamus(UrciteAno) $>:<$ ProcPrace(UrciteAno)
- č. 9: ZdrojeGaudeamus(UrciteAno) $>:<$ ZdrojeDenOtevrenychDveri(UrciteAno)
- č. 10: ZdrojeTelevizeNoviny(SpiseAno) $>:<$ ZdrojeInternetPrf(SpiseAno)

První pravidlo říká, že jestliže se studenti rozhodli studovat kvůli snadnému získání VŠ diplomu, dozvěděli se o možnosti studia od svých spolužáků a známých. Ti mohli studentovi říct, že studium je snadné a lehce se tak dostane k VŠ diplomu. Čtvrté pravidlo ukazuje fakt, že rodiče vyvíjí tlak na své děti,

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	6 %	4 %	6 %
3	16 %	6 %	19 %
4	32 %	8 %	32 %
5	32 %	9 %	32 %
6	17 %	8 %	20 %

Tabulka 14.: Tabulka proc. podílů shluků jednotlivých vah u complete linkage třetí kategorie.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	20 %	90 %	78 %
3	25 %	71 %	65 %
4	33 %	73 %	67 %
5	31 %	74 %	67 %
6	56 %	74 %	63 %

Tabulka 15.: Tabulka proc. podílů shluků jednotlivých vah u Wardovy metody první kategorie.

aby se v budoucnu podílely na výzkumu na vysoké škole. Páté pravidlo má z uvedených pravidel největší podporu a mohlo by ukazovat na fakt, že studenti, kteří navštívili den otevřených dveří, mají opravdový zájem o obor. Toto pravidlo se také nachází ve výsledcích roku 2010, kde má podporu 0,11 a spolehlivost 0,8. Ve výsledcích roku 2010 je obsaženo pravidlo, kde je v antecedentu *ZdrojeFacebook(SpiseNe)*, a tedy protikladné k šestému pravidlu. Toto pravidlo má podporu 0,2 a spolehlivost 0,79. Pravidlo č. 3 a 10 ukazuje, že studenti se dozvídají o studiu z různých informačních kanálů současně.

Pravidla č. 3, 5 a 10 se také nachází ve výsledcích nad sloučenými daty let 2010 a 2011. Pravidlo č. 3 tam má podporu 0,09 a spolehlivost 0,77. Pravidlo č. 5 tam má podporu 0,13 a spolehlivost 0,75. Pravidlo č. 12 tam má podporu 0,05 a spolehlivost 0,77. Jak je vidět, hodnoty podpory a spolehlivosti jsou v těchto výsledcích téměř stejné, a tedy jsou tyto pravidla stabilní.

Z celkového počtu 147 studentů přišlo 46 studentů z gymnázia, 58 studentů ze střední průmyslové školy a 43 studentů ze střední školy.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 18. Většina z pravidel nemá dle tabulky vysokou hodnotu podpory, a tak uvedená pravidla platí pouze pro relativně malou část studentů.

V druhé úloze jsem do antecedentu zvolil pouze atributy studijní obor a typ

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	90 %	77 %	15 %
3	59 %	83 %	54 %
4	62 %	74 %	70 %
5	65 %	70 %	57 %
6	66 %	80 %	46 %

Tabulka 16.: Tabulka proc. podílů shluků jednotlivých vah u Wardovy metody druhé kategorie.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	4 %	5 %	0,7 %
3	34,51 %	35 %	0,7 %
4	4,23 %	12 %	8 %
5	6,34 %	11 %	8 %
6	0 %	70 %	70 %

Tabulka 17.: Tabulka proc. podílů shluků jednotlivých vah u Wardovy metody třetí kategorie.

střední školy. Délku antecedentu jsem neomezoval, jelikož se mohly vytvářet pouze pravidla s maximální délkou dva. Sukcedent a kvantifikátory jsem nechal nastaveny stejně jako v předchozí úloze. Výsledkem úlohy je celkem 8 pravidel. Níže jsou uvedena zajímavá pravidla:

č. 11: StudijniObor(Aplikovaná informatika) & TypStredni(gymnazium) >:< RozhodnutiExkurze(UrciteNe)

č. 12: StudijniObor(Aplikovaná informatika) & TypStredni(gymnazium) >:< RozhodnutiVyzkum(SpiseAno)

č. 13: StudijniObor(Informatika) & TypStredni(gymnazium) >:< ZdrojeFacebook(UrciteNe)

č. 14: StudijniObor(Informatika) & TypStredni(gymnazium) >:< ZdrojePrezentaceNaSs(UrciteNe)

č. 15: TypStredni(ss) >:< ZdrojePrezentaceNaSs(UrciteNe)

Pravidlo č. 12 říká, že studenty z oboru Aplikované informatiky, kteří přišli z gymnázia, ovlivnilo rozhodnutí podílet se na špičkovém výzkumu. U tohoto pravidla bych spíše čekal studenty z oboru Informatika než z oboru Aplikovaná informatika. Pravidlo č. 13 může označovat fakt, že neprobíhala žádná kampaň na Facebooku či jiné sociální síti, které jsou mezi studenty oblíbené. Kampaň

Číslo pravidla	Podpora	Spolehlivost
1	0,08	0,73
2	0,07	0,71
3	0,09	0,71
4	0,12	0,75
5	0,14	0,73
6	0,09	0,71
7	0,17	0,71
8	0,07	0,83
9	0,06	0,75
10	0,06	0,8

Tabulka 18.: Tabulka hodnot podpory a spolehlivosti.

na těchto sítích by mohla zaujmout další studenty ke studiu. Pravidlo č. 15 říká, že studenti, kteří přišli ze středních škol, se nedozvěděli o možnosti studia z prezentací na středních školách. To potvrzuje fakt, že prezentace na těchto oborech probíhaly pouze na středních průmyslových školách.

Jediné pravidlo, které se nachází i v jiných výsledcích je pravidlo č. 13. To se nachází také ve výsledcích sloučených let 2010 a 2011, kde má podobné hodnoty podpory resp. spolehlivosti (0,12 resp. 0,81).

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 19. Hodnoty podpory jsou u uvedených pravidel vyšší než v předchozím případě, přesto žádné z pravidel nepřesahuje podporu 0,25.

Číslo pravidla	Podpora	Spolehlivost
11	0,12	0,71
12	0,12	0,71
13	0,13	0,86
14	0,11	0,71
15	0,22	0,72

Tabulka 19.: Tabulka hodnot podpory a spolehlivosti.

Třetí úlohu tvoří tři podúlohy. Každá z nich tvoří úlohu, ve které jsou kvantifikátory nastaveny stejně jako v předchozích úlohách. V antecedentu i sukcedentu jsou však obsaženy pouze atributy z jednoho okruhu otázek dotazníku. Z vygenerovaných pravidel se mi však žádné nejevilo jako zajímavé. To může být

způsobeno tím, že právě zajímavá pravidla obsahují atributy z různých kategorií otázek dotazníku.

Čtvrtou a pátou úlohu jsem provedl analogicky jako první a druhou úlohu. Data však tvořily pouze záznamy tzv. jader shluků u hierarchického shlukování (celkem 120 objektů). Pravidla ze čtvrté úlohy (celkem 44 pravidel) nevytvořily zajímavý výsledek. Pátou úlohu tvoří celkem 53 pravidel, z nichž dvě pravidla mi přišla zajímavá:

č. 16: TypStredni(gymnazium) >:< RozhodnutiUcitele(UrciteNe)

č. 17: TypStredni(ss) >:< ZdrojePrezentaceNaSs(UrciteNe)

Pravidlo č. 16 říká, že studenty z gymnázií neovlivnili v rozhodování jejich učitelé. Následující pravidlo se již objevilo v předchozí úloze.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 20. Hodnota podpory těchto pravidel je rovna 0,25 (resp. 0,24), tzn. 30 studentů (resp. 29).

Číslo pravidla	Podpora	Spolehlivost
16	0,25	0,73
17	0,24	0,75

Tabulka 20.: Tabulka hodnot podpory a spolehlivosti.

Všechny uvedené úlohy jsem také zkoušel provést s kvantifikátorem dvojitá fundovaná implikace. Výsledky těchto pozměněných úloh obecně obsahovaly méně vygenerovaných pravidel. Navíc žádné z těchto pravidel nebylo nové, ve smyslu již vygenerovaných pravidel fundované implikace.

4.4. Rok 2010 - Apl. informatika a Informatika

Data z oborů Aplikovaná informatika a Informatika z dotazníků roku 2010 obsahují celkem 79 záznamů.

4.4.1. Hierarchické shlukování

Na data jsem aplikoval čtyři metody shlukování. Stejně jako v předchozí úloze shlukování vytvořila metoda complete linkage viz 13. a Wardova metoda viz 14. vhodné dendrogramy pro další zkoumání. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

U metody complete linkage jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 32 objektů, druhý 25 objektů, třetí 19 objektů a čtvrtý shluk 3 objekty. Typičtí zástupci shluků jsou uvedeni v tabulce 48.

První shluk obsahuje část studentů, které ovlivnily v rozhodování exkurze na pracovišti. O možnosti studia se většina dozvěděla z veletrhu Gaudeamus

a brožury PřF UP. Polovina studentů tohoto shluku by se chtěla dát na vědeckou dráhu.

Druhý shluk obsahuje studenty, které v rozhodování ovlivnily nízké životní náklady v Olomouci. O možnosti studia se většina nedozvěděla ze dne otevřených dveří.

Třetí shluk obsahuje studenty, které v rozhodování neovlivnili kamarádi, možnost odstěhovat se od rodičů a být nezávislý, možnost podílet se na výzkumu a nízké životní náklady v Olomouci. Naopak je ovlivnila snadnost přijetí na PřF UP. O možnosti studia se nedozvěděli od spolužáků a známých.

Čtvrtý shluk obsahuje studenty, které ovlivnili v rozhodování rodiče, učitelé na SŠ nebo ZŠ, osobní kontakt se zaměstnancem nebo studentem PřF UP, propagační materiály, exkurze na pracovišti, snadnost přijetí na PřF UP, možnost podílet se na výzkumu a také fakt, že nebyli přijati na jinou VŠ. O možnosti studia se spíše dozvěděli od učitelů na SŠ, z novin a televize, z Facebooku či jiných sociálních sítí a z prezentací na SŠ. Studenti tohoto shluku by se chtěli dát na vědeckou dráhu a myslí si, že je snadné získat VŠ diplom.

Na dendrogramu Wardovy metody jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 24 objektů, druhý 22 objektů, třetí 13 objektů a čtvrtý shluk 20 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 49.

První shluk obsahuje studenty, které ovlivnily v rozhodování propagační materiály, exkurze na pracovišti a možnost podílet se na výzkumu. O možnosti studia se dozvěděli ze dne otevřených dveří. Naopak se o možnosti studia nedozvěděli z veletrhu Gaudeamus. Část studentů se rozhodla studovat pro pozdější vědeckou dráhu.

Druhý shluk obsahuje část studentů, které v rozhodování ovlivnila možnost zůstat v místě bydliště, jejich kamarádi a snadnost přijetí na PřF UP. O možnosti studia se nedozvěděli od učitelů na SŠ a ze dne otevřených dveří.

Třetí shluk obsahuje studenty, které v rozhodování neovlivnily propagační materiály a jejich kamarádi. Naopak je ovlivnila možnost odstěhovat se od rodičů a nízké životní náklady v Olomouci. O možnosti studia se nedozvěděli z brožury PřF UP.

Ve čtvrtém shluku se nachází studenti, které v rozhodování neovlivnili jejich kamarádi. O možnosti studia se nedozvěděli od spolužáků a známých, učitelů na SŠ, z veletrhu Gaudeamus a z brožury PřF UP. Část studentů se rozhodla studovat pro pozdější vědeckou dráhu.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 21. a 22. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Procentuální podíly jsou u metody complete linkage celkově menší než u oboru Aplikovaná informatika a Informatika v roce 2011. U Wardovy metody jsou podíly víceméně vyrovnané.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	23 %	29 %	5 %
3	57 %	58 %	1 %
4	57 %	52 %	18 %
5	34 %	35 %	31 %
6	34 %	35 %	32 %

Tabulka 21.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	5 %	4 %	9 %
3	43 %	14 %	53 %
4	18 %	66 %	65 %
5	57 %	70 %	71 %
6	48 %	81 %	57 %

Tabulka 22.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

4.4.2. Metoda GUHA

Nastavení metody GUHA u první úlohy bylo totožné jako u dat z roku 2011 s obory Aplikovaná informatika a Informatika. Jako kvantifikátor jsem nastavil BASE s hodnotou 10 % a fundovanou implikaci s hodnotou 0,75. Takto zadaná úloha vygenerovala celkem 292 pravidel.

Výběr zajímavých pravidel:

- č. 1: ProcDrahaVaV(UrciteNe) >:< RozhodnutiAtraktivniObor(UrciteAno)
- č. 2: RozhodnutiAtraktivniObor(SpiseAno) >:< RozhodnutiNenarocneStudium(SpiseNe)
- č. 3: RozhodnutiBezPrijimacek(SpiseNe) >:< RozhodnutiPrestizPrf(SpiseAno)
- č. 4: RozhodnutiExkurze(SpiseAno) >:< RozhodnutiPrestizPrf(SpiseAno)
- č. 5: RozhodnutiKamaradi(SpiseNe) >:< ProcLukrativniZamestnani(SpiseAno)
- č. 6: RozhodnutiKamaradi(UrciteNe) >:< ProcZajemObor(UrciteAno)
- č. 7: RozhodnutiPropagacniMaterialy(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 8: ZdrojeFacebook(SpiseAno) >:< RozhodnutiVyzkum(SpiseAno)

Druhé pravidlo má z uvedených pravidel největší podporu. Studenty, které

ovlivnil atraktivní studijní obor, neovlivnila vize nenáročného studia. U těchto studentů tedy vítězí atraktivnost oboru nad nenáročným studiem, což je jistě správně. Čtvrté pravidlo říká, že studenti, kteří se rozhodli dle exkurze na pracovišti, se také rozhodli dle prestiže PřF UP. Je možné, že tito studenti po exkurzi získali dojem o prestiži PřF UP. Páté a šesté pravidlo ukazuje, že studenti, které neovlivnilo rozhodnutí být se svými kamarády, se rozhodli studovat pro získání lukrativního zaměstnání a mají zájem o obor. Tito studenti zřejmě dávají přednost budoucímu získání lukrativního zaměstnání před možností zůstat se svými kamarády. Poslední osmé pravidlo říká, že studenty, kteří se o možnosti studia dozvěděli z Facebooku či jiné sociální sítě, ovlivnilo rozhodnutí podílet se na špičkovém výzkumu. Je možné, že tito studenti viděli výsledky výzkumů na PřF UP na sociálních sítích a mají zájem se na výzkumu podílet.

Až na sedmé pravidlo není žádné pravidlo obsaženo ve výsledcích metody GUHA v roce 2011 a sloučených let 2010 a 2011. Tyto pravidla tak nedosahují minimální hodnoty podpory a spolehlivosti v ostatních datech. Nejsou tedy tak významná a můžeme je označit za zajímavá pouze v aktuálních datech.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 23. Nejmenší podpora pravidel je 0,11, což představuje 8 záznamů. Naopak největší podpora je 0,31, což představuje 31 záznamů.

Číslo pravidla	Podpora	Spolehlivost
1	0,16	0,8
2	0,31	0,77
3	0,15	0,86
4	0,18	0,78
5	0,20	0,75
6	0,23	0,77
7	0,11	0,89
8	0,11	0,8

Tabulka 23.: Tabulka hodnot podpory a spolehlivosti.

V druhé úloze jsem do antecedentu zvolil pouze atributy studijní obor a typ střední školy. Délku antecedentu jsem neomezoval, jelikož se mohly vytvářet pouze pravidla s maximální délkou dva. Sukcedent a kvantifikátory jsem nechal nastaveny stejně jako v předchozí úloze. Výsledkem úlohy je celkem 47 pravidel. Níže jsou uvedena zajímavá pravidla:

č. 9: StudijniObor(Aplikovaná informatika) & TypStredni(ss) >:<

ProcLukrativniZamestnani(SpiseAno)

č. 10: TypStredni(gymnazium) >:< ProcPrace(UrciteAno)

č. 11: TypStredni(gymnazium) >:< ProcZajemObor(UrciteAno)

Pravidlo č. 9 říká, že studenti z oboru Aplikovaná informatika, kteří přišli ze střední školy, se rozhodli studovat kvůli lukrativnímu zaměstnání. Tito studenti ze střední školy zřejmě přicházejí studovat tento obor s více aplikovanými předměty s vidinou budoucího lukrativního zaměstnání. Další dvě pravidla uvádějí, že studenti z gymnázií se rozhodli studovat kvůli práci, pro kterou je obor potřeba, a protože mají zájem o obor, což je jistě správné.

Žádné z uvedených pravidel se nenachází ve výsledcích dat z roku 2011 nebo sloučených let 2010 a 2011. Nalezená pravidla tak nemají v ostatních datech dostatečnou hodnotu podpory a spolehlivosti a jsou platná pouze v aktuálních datech.

Z celkového počtu 79 studentů přišlo 15 studentů z gymnázia, 35 studentů ze střední průmyslové školy a 28 studentů ze střední školy.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 24.

Číslo pravidla	Podpora	Spolehlivost
24	0,13	0,83
25	0,16	0,86
26	0,16	0,86

Tabulka 24.: Tabulka hodnot podpory a spolehlivosti.

4.5. Rok 2010 a 2011 - Apl. informatika a Informatika

Data z oborů Aplikovaná informatika a Informatika z dotazníků roku 2010 a 2011 obsahují celkem 226 záznamů.

4.5.1. Hierarchické shlukování

Na data jsem aplikoval čtyři metody shlukování. Stejně jako v předchozí úloze shlukování vytvořila metoda complete linkage viz 15. a Wardova metoda viz 16. vhodné dendrogramy pro další zkoumání. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

Hladinu se čtyřmi shluky jsem zvolil na metodě complete linkage. První shluk obsahuje 51 objektů, druhý 133 objektů, třetí 37 objektů a čtvrtý shluk 5 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 50.

První shluk obsahuje většinu studentů, kteří byli při rozhodování ovlivněni osobním kontaktem se zaměstnancem nebo studentem PŘF UP, exkurzí na pracovišti, kamarády a možností zůstat v místě bydliště.

Část studentů ve druhém shluku byla při rozhodování ovlivněna nenáročností studia. Značná část studentů se v tomto shluku nedozvěděla o možnosti studia z internetových stránek PřF UP a ze dne otevřených dveří. Část studentů se však dozvěděla o studiu z Facebooku či jiné sociální sítě. Tento shluk také obsahuje skupinu studentů, kteří se rozhodli studovat pro budoucí vědeckou dráhu.

V třetím shluku se nachází studenti, kteří při rozhodování byli převážně ovlivněni možností podílet se na výzkumu. O studiu se dozvěděli převážně od spolužáků a známých, učitelů na SŠ, z veletrhu Gaudeamus a brožury PřF UP. Téměř polovina studentů se rozhodla studovat pro snadné získání VŠ diplomu.

Čtvrtý shluk obsahuje studenty, kteří byli při rozhodování ovlivněni snadností přijetí na PřF UP. Nebyli naopak ovlivněni možností podílet se na výzkumu.

U Wardovy metody jsem zvolil hladinu se třemi shluky. První shluk obsahuje 124 objektů, druhý 52 objektů a třetí shluk 50 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 51.

První shluk obsahuje část studentů, kteří byli při rozhodování ovlivněni rodiči, propagačními materiály a nízkými životními náklady v Olomouci. O možnosti studia se dozvěděli převážně od spolužáků a známých. Naopak se nedozvěděli z brožury PřF UP.

Studenti ve druhém shluku byli při rozhodování ovlivněni propagačními materiály, exkurzí, možností zůstat v místě bydliště, výzkumem a nízkými životními náklady v Olomouci. O možnosti studia se dozvěděli převážně od spolužáků a známých a brožury PřF UP. Část studentů se rozhodla studovat kvůli snadnému VŠ diplomu.

Ve třetím shluku se nachází studenti, kteří nebyli ovlivněni při rozhodování učiteli na SŠ a osobním kontaktem. Naopak byli ovlivněni možností zůstat v místě bydliště. O možnosti studia se nedozvěděli od spolužáků a známých, učitelů na SŠ, ze dne otevřených dveří, z veletrhu Gaudeamus a brožury PřF UP.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 25. a 26. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Procentuální podíly jsou u metody complete linkage většinou mírně menší než podíly v datech jednotlivých let. U Wardovy metody jsou podíly mezi vahami #1 a #3 kromě případu dvou shluků menší. Mezi ostatními vahami jsou podíly mírně větší.

4.5.2. Metoda GUHA

Nastavení první úlohy je stejné jako první úlohy roku 2011. Jako kvantifikátor jsem nastavil BASE s absolutní hodnotou 10 (nejméně 10 záznamů musí splňovat předpoklad i závěr) a fundovanou implikaci s hodnotou 0,75. Takto zadaná úloha vygenerovala celkem 233 pravidel.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	73 %	17 %	62 %
3	79 %	30 %	62 %
4	76 %	30 %	63 %
5	71 %	85 %	72 %
6	74 %	84 %	74 %

Tabulka 25.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	7 %	87 %	83 %
3	54 %	88 %	57 %
4	34 %	58 %	48 %
5	79 %	77 %	50 %
6	71 %	79 %	50 %

Tabulka 26.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

Výběr zajímavých pravidel:

- č. 1: ProcDrahaVaV(UrciteAno) >:< ProcLukrativniZamestnani(UrciteAno)
- č. 2: ProcDrahaVaV(UrciteAno) >:< ProcPrace(UrciteAno)
- č. 3: RozhodnutiNizkeZivotniNaklady(UrciteAno) >:< RozhodnutiAtraktivniObor(UrciteAno)
- č. 4: RozhodnutiOdstehovani(UrciteAno) >:< ProcPrace(UrciteAno)
- č. 5: RozhodnutiPropagacniMaterialy(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 6: RozhodnutiRodice(SpiseAno) >:< RozhodnutiSpolupraceProjekty(UrciteNe)
- č. 7: ZdrojeBrozuraPrf(UrciteAno) >:< ProcLukrativniZamestnani(UrciteAno)
- č. 8: ZdrojeTelevizeNoviny(SpiseAno) >:< ZdrojeInternetPrf(SpiseAno)
- č. 9: ZdrojeUciteleSs(UrciteAno) >:< ProcPrace(UrciteAno)

První a druhé pravidlo říká, že studenti, kteří se rozhodli studovat pro vědeckou dráhu a pozdější výzkum v této oblasti, se rozhodli studovat také pro získání lukrativního zaměstnání a práci, pro níž je obor potřeba. U těchto pravidel získávám dojem, že studenti při nástupu na studium zatím netuší, jestli se chtějí

později vydat na vědeckou dráhu nebo získat lukrativní práci v tomto oboru. Čtvrté pravidlo uvádí, že studenti, které ovlivnilo rozhodnutí možnost odstěhovat se od rodičů, se rozhodli studovat kvůli práci, pro níž je tento obor potřeba. Tito studenti tak mají zájem o pozdější práci v oboru a neváhají se proto odstěhovat od rodičů. Poslední pravidlo říká, že studenti, kteří se dozvěděli o studiu od učitelů na střední škole, se rozhodli studovat kvůli práci, pro níž je tento obor potřeba. Je možné, že tito studenti jsou motivováni učiteli středních škol, kteří tvrdí, že po absolvování tohoto oboru by měli mít větší pracovní uplatnění.

Třetí a páté pravidlo se také nachází v datech z roku 2010. Třetí pravidlo tam má podporu 0,12 a spolehlivost 0,82; a tedy mírně vyšší než v aktuálních datech. Páté pravidlo tam má podporu 0,11 a spolehlivost 0,89. Pravidla č. 1 (podpora 0,06 a spolehlivost 0,89), 3 (podpora 0,06 a spolehlivost 0,75), 5 (podpora 0,09 a spolehlivost 0,71) a 8 (podpora 0,06 a spolehlivost 0,8) jsou obsažena ve výsledcích roku 2011. Hodnoty podpory a spolehlivosti jsou u uvedených pravidel téměř stejné jako u zde nalezených pravidel.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 27. Nejmenší hodnota podpory uvedených pravidel je 0,05, což představuje 11 záznamů. Největší hodnota podpory je 0,11, což je 24 záznamů.

Číslo pravidla	Podpora	Spolehlivost
1	0,06	0,86
2	0,05	0,79
3	0,08	0,78
4	0,06	0,75
5	0,09	0,77
6	0,11	0,77
7	0,06	0,76
8	0,05	0,77
9	0,06	0,8

Tabulka 27.: Tabulka hodnot podpory a spolehlivosti.

V druhé úloze jsem do antecedentu zvolil pouze atributy studijní obor a typ střední školy. Délku antecedentu jsem neomezoval, jelikož se mohly vytvářet pouze pravidla s maximální délkou dva. Sukcedent a kvantifikátory jsem ponechal nastaveny stejně jako v předchozí úloze. Výsledkem úlohy je celkem 53 pravidel. Níže je uvedeno zajímavé pravidlo:

č. 10: StudijniObor(Informatika) & TypStredni(gymnazium) >:< ProcZajemObor(UrciteAno)

Toto pravidlo se objevilo také ve výsledcích roku 2010 a i 2011. V roce 2010 mělo podporu 0,12 a spolehlivost 0,9. V roce 2011 mělo podporu 0,12 a spolehlivost 0,81. Pravidlo je tak ve všech datech stabilní.

Z celkového počtu 226 studentů přišlo 61 studentů z gymnázia, 92 studentů ze střední průmyslové školy a 68 studentů ze střední školy.

Hodnoty podpory a spolehlivosti pravidla jsou uvedeny v tabulce 28. Podpora 0,12 představuje 26 záznamů.

Číslo pravidla	Podpora	Spolehlivost
10	0,12	0,84

Tabulka 28.: Tabulka hodnot podpory a spolehlivosti.

4.6. Rok 2011 - všechny obory

Data z roku 2011 obsahují celkem 1124 záznamů.

4.6.1. Hierarchické shlukování

Ze vstupní datové matice jsem na počátku shlukování odstranil řádky, kde počet neznámých hodnot odpovědí byl větší než 16. Při neodstranění těchto řádků nebylo možné vypočítat matici vzdáleností pomocí funkce *dist*.

Na data jsem aplikoval čtyři metody shlukování. Stejně jako v předchozí úloze shlukování vytvořila metoda complete linkage a Wardova metoda vhodné dendrogramy pro další zkoumání viz 17. resp. 18. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

U metody complete linkage jsem zvolil hladinu se třemi shluky. První shluk obsahuje 395 objektů, druhý 594 objektů a třetí shluk 108 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 52.

První shluk obsahuje studenty, které při rozhodování ovlivnily propagační materiály a možnost výzkumu. O možnosti studia se dozvěděli od spolužáků a známých.

Druhý shluk obsahuje studenty, které při rozhodování ovlivnil osobní kontakt se zaměstnancem nebo studentem PřF UP, propagační materiály, exkurze na pracovišti, možnost výzkumu a nízké životní náklady v Olomouci. Část studentů také ovlivnila spolupráce na projektech. Polovinu studentů ovlivnili při rozhodování kamarádi. O možnosti studia se dozvěděli převážně od spolužáků a známých, ze dne otevřených dveří, veletrhu Gaudeamus. Část studentů se o studiu dozvěděla z prezentací na SŠ. Téměř polovina studentů se rozhodla studovat pro pozdější vědeckou dráhu.

V třetím shluku se nachází studenti, které při rozhodování převážně ovlivnila snadnost přijetí na PřF UP. O možnosti studia se nedozvěděli ze dne otevřených dveří a brožury PřF UP. Polovina studentů se nerozhodla studovat kvůli lukrativnímu zaměstnání.

Při Wardově metodě jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 378 objektů, druhý 94 objektů, třetí 200 objektů a čtvrtý shluk 425 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 53.

V prvním shluku se nachází studenti, které převážně ovlivnily při rozhodování propagační materiály a snadnost přijetí na PřF UP. O možnosti studia se dozvěděli od spolužáků a známých.

Druhý shluk obsahuje studenty, které neovlivnila možnost výzkumu. O možnosti studia se nedozvěděli od učitelů na SŠ, ze dne otevřených dveří a brožury PřF UP.

Třetí shluk obsahuje studenty, které při rozhodování neovlivnili učitelé. Naopak je ovlivnili kamarádi.

Ve čtvrtém shluku jsou studenti, které při rozhodování převážně ovlivnil osobní kontakt se zaměstnancem nebo studentem PřF UP, propagační materiály, exkurze na pracovišti, možnost výzkumu a nízké životní náklady v Olomouci. O možnosti studia se dozvěděli od spolužáků a známých, ze dne otevřených dveří a veletrhu Gaudeamus. Část studentů se dozvěděla o studiu z Facebooku či jiné sociální sítě a z prezentací na SŠ. Většina studentů se rozhodla studovat pro pozdější vědeckou dráhu.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 29. a 30. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	18 %	83 %	87 %
3	75 %	50 %	84 %
4	84 %	51 %	86 %
5	77 %	72 %	85 %
6	83 %	73 %	92 %

Tabulka 29.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Procentuální podíly jsou u obou metod většinou vyšší než u dat z oboru Aplikovaná informatika a Informatika z roku 2010.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	27 %	22 %	21 %
3	51 %	29 %	46 %
4	61 %	33 %	58 %
5	63 %	49 %	54 %
6	81 %	54 %	80 %

Tabulka 30.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

Data s obory

Předchozí data obsahovala pro shlukování pouze všechny otázky z dotazníků. Připravil jsem variantu, kdy jsem do dat pro shlukování přidal i informaci o oboru studenta. Data oborů jsem naškáloval tak, že každý obor má svůj sloupec. U každého záznamu (řádku) je hodnota 1 v tom sloupci, který odpovídá oboru záznamu. V ostatních sloupcích oborů jsou hodnoty 0.

Na tato data jsem opět aplikoval čtyři metody shlukování. Stejně jako v předchozí úloze shlukování vytvořila metoda complete linkage viz 19. a Wardova metoda viz 20. vhodné dendrogramy pro další zkoumání. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

Typičtí zástupci jsou vybráni z datové matice, kde byly odstraněny ty řádky, kde počet neznámých hodnot odpovědí byl větší než 16. Při neodstranění těchto řádků se mezi typickými zástupci objevovali zástupci s příliš mnoha neznámými hodnotami.

U metody complete linkage jsem zvolil hladinu se třemi shluky. První shluk obsahuje 593 objektů, druhý 510 objektů a třetí shluk 21 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 54.

V prvním shluku se nachází studenti, které při rozhodování převážně ovlivnila snadnost přijetí na PřF UP a možnost výzkumu. O možnosti studia se dozvěděli z internetových stránek PřF UP. V tomto shluku se nachází 44 z celkem 56 studentů oboru Informatika. Z oboru Aplikovaná informatika je v tomto shluku 67 z celkem 91 studentů.

Druhý shluk obsahuje studenty, které při rozhodování převážně ovlivnil osobní kontakt se zaměstnancem nebo studentem PřF UP a propagační materiály. O možnosti studia se dozvěděli od spolužáků a známých, učitelů na SŠ, z internetových stránek PřF UP, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Část studentů se také dozvěděla o studiu z prezentací na SŠ. Studenti se rozhodli studovat kvůli budoucí vědecké dráze.

Ve třetím shluku jsou studenti, které převážně při rozhodování ovlivnila snadnost přijetí na PřF UP, možnost výzkumu, nízké životní náklady v Olomouci

a vize nenáročného studia. O možnosti studia se dozvěděli od spolužáků a známých a z Facebooku či jiné sociální sítě. Studenti se rozhodli studovat kvůli budoucí vědecké dráze. Většina se také rozhodla studovat pro vidinu snadného získání VŠ diplomu.

U Wardovy metody jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 342 objektů, druhý 399 objektů, třetí 65 objektů a čtvrtý shluk 318 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 55.

První shluk obsahuje studenty, kteří byli při rozhodování ovlivněni propagačními materiály a možností podílet se na výzkumu. O možnosti studia se dozvěděli od spolužáků a známých. V tomto shluku se nachází 50 z celkem 56 studentů oboru Informatika.

V druhém shluku jsou studenti, kteří byli při rozhodování ovlivněni propagačními materiály, exkurzí na pracovišti, možností podílet se na výzkumu a nízkými životními náklady v Olomouci. O možnosti studia se dozvěděli od spolužáků a známých, učitelů na SŠ, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Část studentů se dozvěděla o možnosti studia z televize a novin. Většina studentů se rozhodla studovat pro budoucí vědeckou dráhu.

Studenti třetího shluku nebyli při rozhodování ovlivněni svými rodiči, učiteli na SŠ a propagačními materiály. Více než polovina studentů se nerozhodla podle budoucího lukrativního zaměstnání a práci, pro kterou je příslušný obor potřeba.

Čtvrtý shluk obsahuje studenty, kteří byli při rozhodování ovlivněni propagačními materiály, možností zůstat v místě bydliště, snadností přijetí na PřF UP a nízkými životními náklady v Olomouci. O možnosti studia se dozvěděli od spolužáků a známých. Z oboru Aplikovaná informatika je v tomto shluku 69 z celkem 91 studentů.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 31. a 32. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	64 %	23 %	65 %
3	71 %	32 %	76 %
4	72 %	30 %	76 %
5	82 %	53 %	78 %
6	77 %	76 %	56 %

Tabulka 31.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	22 %	20 %	20 %
3	31 %	30 %	27 %
4	46 %	49 %	27 %
5	69 %	78 %	30 %
6	67 %	72 %	34 %

Tabulka 32.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

Procentuální podíly jsou u obou metod velice podobné shlukování bez sloupců s obory. Podíly mezi vahami #1 a #3 jsou však zde nižší.

4.6.2. Metoda GUHA

Nastavení první úlohy je totožné jako nastavení úlohy roku 2011 s obory Aplikovaná informatika a Informatika. Jako kvantifikátor jsem nastavil BASE s absolutní hodnotou 10 a fundovanou implikaci s hodnotou 0,75. Takto zadaná úloha vygenerovala celkem 323 pravidel.

Výběr zajímavých pravidel:

- č. 1: ProcZajemObor(SpiseNe) >:< Kraj(Olomoucký)
- č. 2: ProcZajemObor(UrciteNe) >:< RozhodnutiRodice(UrciteNe)
- č. 3: RozhodnutiPropagacniMaterialy(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 4: RozhodnutiUcitele(UrciteAno) >:< ProcZajemObor(UrciteAno)
- č. 5: StudijniObor(Bioanorganická chemie) >:< ZdrojeInternetPrf(UrciteAno)
- č. 6: StudijniObor(Biologie - zeměpis) >:< ProcDrahaVaV(SpiseNe)
- č. 7: ZdrojeSoutezeSoustredeni(UrciteAno) >:< ProcZajemObor(UrciteAno)
- č. 8: ZdrojeTelevizeNoviny(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 9: Kraj(Žilinský) >:< RozhodnutiAtraktivniObor(SpiseAno)
- č. 10: ProcSnadnyDiplom(UrciteAno) >:< Kraj(Olomoucký)
- č. 11: RozhodnutiSpolupraceProjekty(UrciteAno) >:< ProcSnadnyDiplom(UrciteNe)
- č. 12: StudijniObor(Optometrie) >:< ProcPrace(UrciteAno)
- č. 13: ZdrojeFacebook(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 14: ZdrojePrezentaceNaSs(UrciteAno) >:< ProcZajemObor(UrciteAno)

První pravidlo říká, že studenti, kteří se nerozhodli studovat, protože je obor baví a zajímá, jsou z Olomouckého kraje. Toto pravidlo má spolehlivost 0,86, což je 30 studentů z 35, kteří splňují pouze předpoklad. Třetí pravidlo dosahuje

z uvedených pravidel nejvyšší podpory. Vyplyvá z něj, že studenti, kteří byli ovlivněni při rozhodování propagačními materiály, se dozvěděli o možnosti studia z internetových stránek PřF UP. Pravidlo č. 5 poukazuje na to, že studenti oboru Bioanorganická chemie se dozvěděli o studiu z internetových stránek PřF UP. Další pravidlo říká, že studenti oboru Biologie - zeměpis se nerozhodovali podle budoucí dráhy vědeckého pracovníka. Pravidlo č. 9 ukazuje, že studenty z Žilinského kraje na Slovensku ovlivnila atraktivnost oboru. Desáté pravidlo uvádí, že studenti, kteří rozhodli studovat kvůli snadné cestě získání VŠ diplomu, pochází z Olomouckého kraje. Stejně jako u prvního pravidla je toto pravidlo zajímavé a zřejmě chtějí tito studenti co nejrychleji získat VŠ diplom, aniž by projevili o obor zájem. Pravidlo č. 12 říká, že studenti oboru Optometrie se rozhodli studovat kvůli práci, pro níž je tento obor potřeba.

Prvních osm pravidel se nachází také ve výsledcích z dat sloučených let 2010 a 2011. Hodnoty podpory a spolehlivosti v těchto výsledcích jsou uvedeny v tabulce 33. ve sloupci Podpora 2 a Spolehlivost 2. Pravidla č. 3, 4 a 8 se nacházejí ve výsledcích z roku 2010. Pravidlo č. 3 zde má podporu 0,10 a spolehlivost 0,77; pravidlo č. 4 má podporu 0,03 a spolehlivost 0,83 a pravidlo č. 8 má podporu 0,02 a spolehlivost 0,85. Tyto tři pravidla bych považoval za stabilní, jelikož se vyskytují ve všech výsledcích dat. Všechna tato společná pravidla mají velmi podobné hodnoty podpory a spolehlivosti.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 33. Všechna pravidla kromě třetího mají malou podporu, která odpovídá desítkám záznamů.

4.7. Rok 2010 - všechny obory

Data z roku 2010 obsahují celkem 577 záznamů ze všech oborů.

4.7.1. Hierarchické shlukování

Ze vstupní datové matice jsem na počátku shlukování odstranil řádky, kde počet neznámých hodnot odpovědí byl větší než 16. Při neodstranění těchto řádků nebylo možné vypočítat matici vzdáleností.

Na data jsem opět aplikoval čtyři metody shlukování. Metody complete linkage viz 21. a Wardova metoda viz 22. vytvořily vhodné dendrogramy pro další zkoumání. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

U metody complete linkage jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 201 objektů, druhý 333 objektů, třetí shluk 20 objektů a čtvrtý shluk 13 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 56.

V prvním shluku jsou studenti, které při rozhodování ovlivnily propagační materiály, exkurze na pracovišti PřF UP a možnost podílet se na výzkumu. O možnosti studia se dozvěděli především od učitelů na SŠ.

Č. prav.	Podpora	Spolehlivost	Podp. 2	Spolehl. 2
1	0,03	0,86	0,02	0,82
2	0,01	0,87	0,01	0,83
3	0,13	0,82	0,12	0,80
4	0,03	0,85	0,03	0,86
5	0,01	0,79	0,01	0,79
6	0,01	0,78	0,01	0,78
7	0,01	0,76	0,01	0,80
8	0,02	0,75	0,00	0,78
9	0,01	0,85	-	-
10	0,01	0,85	-	-
11	0,02	0,75	-	-
12	0,02	0,76	-	-
13	0,03	0,78	-	-
14	0,03	0,75	-	-

Tabulka 33.: Tabulka hodnot podpory a spolehlivosti.

Druhý shluk obsahuje studenty, které při rozhodování ovlivnily propagační materiály a možnost podílet se na výzkumu. O možnosti studia se dozvěděli především od spolužáků a známých. Naopak zde existuje skupina studentů, kteří se o studiu nedozvěděli z internetových stránek PřF UP, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP.

Třetí shluk obsahuje studenty, které při rozhodování ovlivnil osobní kontakt se zaměstnancem nebo studentem PřF UP, spolupráce na projektech PřF UP během studia na SŠ, propagační materiály, možnost zůstat se svými kamarády a podílet se na výzkumu. Část studentů neovlivnila prestiž PřF UP. O možnosti studia se dozvěděli především od spolužáků a známých, učitelů na SŠ a prezentací na SŠ. Většina studentů se nedozvěděla o možnosti studia z veletrhu Gaudeamus. Studenti se rozhodli studovat obor pro lukrativní zaměstnání a vědeckou dráhu.

Ve čtvrtém shluku se nachází studenti, které při rozhodování neovlivnily propagační materiály, exkurze na pracovišti PřF UP, možnost podílet se na výzkumu. Naopak studenty ovlivnila snadnost přijetí na PřF UP a fakt, že nebyli přijati na jinou VŠ. Většina studentů se nedozvěděla o možnosti studia ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Studenti se nerozhodli studovat obor pro lukrativní zaměstnání a vědeckou dráhu.

Vhodná hladina u Wardovy metody tvořila čtyřmi shluky. První shluk obsahuje 183 objektů, druhý 88 objektů, třetí 143 objektů a čtvrtý shluk 153 objektů.

Typičtí zástupci shluků jsou uvedeni v tabulce 57.

První shluk obsahuje studenty, kteří byli při rozhodování ovlivněni propagačními materiály, exkurzí a výzkumem. Naopak nebyli ovlivněni možnostmi odstěhovat se od rodičů. O možnosti studia se dozvěděli především od spolužáků a známých, učitelů na SŠ, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Studenti se rozhodli studovat obor pro vědeckou dráhu.

Ve druhém shluku se nachází studenti, které při rozhodování ovlivnila možnost zůstat v místě bydliště, kamarádi a snadnost přijetí na PřF UP. Naopak nebyli ovlivněni prestiží PřF UP. O možnosti studia se dozvěděli především od spolužáků a známých.

Třetí shluk obsahuje studenty, kteří byli při rozhodování ovlivněni propagačními materiály, exkurzí, možnostmi zůstat v místě bydliště, kamarády, výzkumem a nízkými životními náklady v Olomouci. O možnosti studia se dozvěděli především od spolužáků a známých, ze dne otevřených dveří a brožury PřF UP. Studenti se rozhodli studovat obor pro vědeckou dráhu.

Čtvrtý shluk neobsahuje v žádné otázce výraznější převahu odpovědí.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 34. a 35. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	56 %	59 %	5 %
3	62 %	70 %	23 %
4	56 %	39 %	60 %
5	56 %	64 %	69 %
6	54 %	73 %	80 %

Tabulka 34.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Procentuální podíly jsou u obou metod mírně menší než v případě shlukování z roku 2011. Jsou však zde i výjimky, kdy jsou podíly větší.

Data s obory

Stejně jako na datech z roku 2011 jsem připravil variantu, kdy jsem do dat pro shlukování přidal škálováním i informaci o oboru studenta.

Na tato data jsem opět aplikoval čtyři metody shlukování. Stejně jako v předchozí úloze shlukování vytvořila metoda complete linkage viz 23. a Wardova

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	76 %	17 %	79 %
3	66 %	24 %	68 %
4	55 %	30 %	53 %
5	59 %	52 %	55 %
6	74 %	77 %	59 %

Tabulka 35.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

metoda viz 24. vhodné dendrogramy pro další zkoumání. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

Typičtí zástupci jsou vybráni z datové matice, kde byly odstraněny ty řádky, kde počet neznámých hodnot odpovědi byl větší než 16. Při neodstranění těchto řádků se mezi typickými zástupci objevovali zástupci s příliš mnoha neznámými hodnotami.

U metody complete linkage jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 312 objektů, druhý 210 objektů, třetí shluk 38 objektů a čtvrtý shluk 17 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 58.

První shluk obsahuje studenty, které při rozhodování ovlivnily propagační materiály, exkurze na pracovišti PřF UP, možnost podílet se na výzkumu, nízké životní náklady v Olomouci a prestiž PřF UP. O možnosti studia se dozvěděli především od spolužáků a známých, učitelů na SŠ, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP.

Studenty ve druhém shluku ovlivnila při rozhodování možnost podílet se na výzkumu a prestiž PřF UP. O možnosti studia se dozvěděli především od spolužáků a známých. Část studentů se o možnosti studia nedozvěděla z internetových stránek PřF UP.

Třetí shluk obsahuje studenty, které při rozhodování ovlivnila exkurze na pracovišti PřF UP, možnost zůstat v místě bydliště, podílet se na výzkumu, nízké životní náklady v Olomouci a prestiž PřF UP. O možnosti studia se dozvěděli především od učitelů na SŠ, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP.

Ve čtvrtém shluku se nachází studenti, kteří byli ovlivněni při rozhodování snadností přijetí na PřF UP. Naopak nebyli ovlivněni atraktivností oboru. Studenti se nerozhodli studovat obor pro lukrativní zaměstnání a vědeckou dráhu.

U Wardovy metody jsem zvolil hladinu s pěti shluky. První shluk obsahuje 137 objektů, druhý 88 objektů, třetí 160 objektů, čtvrtý shluk 37 objektů a pátý shluk 155 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 59.

První shluk obsahuje studenty, které při rozhodování ovlivnily propagační ma-

teriály a možností podílet se na výzkumu. O možnosti studia se dozvěděli ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Studenti se rozhodli studovat obor pro vědeckou dráhu.

Studenty ve druhém shluku ovlivnily při rozhodování propagační materiály, exkurze na pracovišti PřF UP, možnost podílet se na výzkumu a nízké životní náklady v Olomouci. O možnosti studia se dozvěděli především od spolužáků a známých, učitelů na SŠ, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Studenti se rozhodli studovat obor pro vědeckou dráhu.

Třetí shluk obsahuje studenty, které při rozhodování ovlivnila snadnost přijetí na PřF UP. O možnosti studia se dozvěděli především od spolužáků a známých.

Ve čtvrtém shluku se nachází studenti, kteří byli ovlivněni při rozhodování propagačními materiály, exkurzí na pracovišti PřF UP, kamarády a možností podílet se na výzkumu. O možnosti studia se dozvěděli především od spolužáků a známých, učitelů na SŠ, ze dne otevřených dveří a veletrhu Gaudeamus. Studenti se rozhodli studovat obor pro vědeckou dráhu. Tento shluk tvoří až na jeden objekt studenti z oboru Regionální geografie.

V pátém shluku jsou studenti, kteří byli při rozhodování ovlivněni propagačními materiály, možností zůstat v místě bydliště, snadností přijetí na PřF UP a nízkými životními náklady v Olomouci. O možnosti studia se dozvěděli především od spolužáků a známých, ze dne otevřených dveří a brožury PřF UP.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 36. a 37. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovídají jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	29 %	35 %	55 %
3	38 %	50 %	65 %
4	36 %	37 %	38 %
5	39 %	40 %	40 %
6	77 %	77 %	52 %

Tabulka 36.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Procentuální podíly jsou metody complete linkage až na pár výjimek mírně menší než v případě shlukování bez sloupců oborů. U Wardovy metody jsou podíly mírně větší.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	33 %	46 %	27 %
3	25 %	34 %	31 %
4	69 %	66 %	60 %
5	73 %	85 %	31 %
6	76 %	71 %	72 %

Tabulka 37.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

4.7.2. Metoda GUHA

Nastavení této první úlohy je opět stejné nastavení první úlohy v roce 2011. Kvantifikátor jsem nastavil BASE s absolutní hodnotou 10 a fundovanou implikací s hodnotou 0,75. Takto zadaná úloha vygenerovala celkem 275 pravidel.

Výběr zajímavých pravidel:

- č. 1: RozhodnutiPropagacniMaterialy(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 2: RozhodnutiUcitele(UrciteAno) >:< ProcZajemObor(UrciteAno)
- č. 3: ZdrojeTelevizeNoviny(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 4: RozhodnutiExkurze(UrciteAno) >:< ZdrojeDenOtevrenychDveri(UrciteAno)
- č. 5: RozhodnutiSpolupraceProjekty(UrciteAno) >:< Kraj(Olomoucký)
- č. 6: StudijniObor(Aplikovaná chemie) >:< ZdrojeInternetPrf(UrciteAno)
- č. 7: StudijniObor(Optometrie) >:< RozhodnutiAtraktivniObor(UrciteAno)
- č. 8: StudijniObor(Optometrie) >:< RozhodnutiPrestizPrf(SpiseAno)
- č. 9: Kraj(Vysočina) >:< RozhodnutiPrestizPrf(SpiseAno)

První pravidlo se nachází i v datech roku 2011, ale s větší podporou. Tato větší podpora v roce 2011 by mohla znamenat vzrůstající tendenci hledání vhodných oborů na Internetu. Druhé pravidlo se taktéž nachází i v datech roku 2011. Třetí pravidlo se také nachází v datech roku 2011 s mírně menší podporou. Pravidlo č. 4 říká, že pokud studenty ovlivnila exkurze nebo akce pořádaná PřF UP, dozvěděli se o možnosti studia ze dne otevřených dveří. Je možné, že tito studenti po absolvování exkurze navštívili také den otevřených dveří, aby o studiu zjistili více informací. Páté pravidlo ukazuje, že studenti, kteří byli ovlivněni spoluprací na projektech na střední škole, jsou z Olomouckého kraje.

První tři pravidla jsou také obsaženy ve výsledcích dat z roku 2011 a sloučených let 2010 a 2011. Hodnoty podpory a spolehlivosti těchto pravidel jsou uvedeny v tabulce 38. Ve sloupci Podpora 2 a Spolehlivost 2 jsou uvedeny hodnoty pro data z roku 2011. Ve sloupci Podpora 3 a Spolehlivost 3 jsou uvedeny

hodnoty pro data z roku 2010 a 2011. Jak je z tabulky vidět, hodnoty podpory a spolehlivosti jsou velmi podobné, a tak tyto pravidla považuji za velmi stabilní.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 38. Všechna uvedená pravidla mají malou hodnotu podpory.

Č. prav.	Podp.	Spolehl.	Podp. 2	Spolehl. 2	Podp. 3	Spolehl. 3
1	0,1	0,77	0,13	0,82	0,12	0,8
2	0,3	0,83	0,03	0,85	0,03	0,85
3	0,2	0,85	0,02	0,75	0,02	0,78
4	0,09	0,75	-	-	-	-
5	0,02	0,79	-	-	-	-
6	0,02	0,75	-	-	-	-
7	0,02	0,83	-	-	-	-
8	0,02	0,77	-	-	-	-
9	0,02	0,92	-	-	-	-

Tabulka 38.: Tabulka hodnot podpory a spolehlivosti.

4.8. Rok 2010 a 2011 - všechny obory

Data z roku 2010 a 2011 obsahují celkem 1701 záznamů ze všech oborů.

4.8.1. Hierarchické shlukování

Ze vstupní datové matice jsem opět na počátku shlukování odstranil řádky, kde počet neznámých hodnot odpovědí byl větší než 16. Při neodstranění těchto řádků nebylo možné vypočítat matici vzdáleností.

Na data jsem opět aplikoval čtyři metody shlukování. Metody complete linkage a Wardova metoda vytvořily vhodné dendrogramy pro další zkoumání viz 25. resp. 26.. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

U metody complete linkage jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 883 objektů, druhý 406 objektů, třetí shluk 324 objektů a čtvrtý shluk 51 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 60.

První shluk obsahuje studenty, které při rozhodování ovlivnily propagační materiály, exkurze na pracovišti PřF UP a možnost podílet se na výzkumu. O možnosti studia se studenti dozvěděli od učitelů na SŠ, ze dne otevřených dveří a veletrhu Gaudeamus. Více než polovina studentů se rozhodla studovat pro pozdější vědeckou dráhu.

Téměř polovinu studentů ve druhém shluku ovlivnila při rozhodování snadnost přijetí na PřF UP. O možnosti studia se studenti nedozvěděli ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP.

Ve třetím shluku se nachází studenti, kteří byli při rozhodování ovlivněni propagačními materiály, možností podílet se na výzkumu a nízkými životními náklady v Olomouci. O možnosti studia se studenti dozvěděli z veletrhu Gaudeamus.

Čtvrtý shluk obsahuje téměř polovinu studentů, které při rozhodování ovlivnili rodiče a dále učitelé na SŠ, osobní kontakt se zaměstnancem nebo studentem PřF UP, propagační materiály, exkurze na pracovišti PřF UP, kamarádi, snadnost přijetí na PřF UP, možnost podílet se na výzkumu, nízké životní náklady v Olomouci a vize nenáročného studia. O možnosti studia se studenti dozvěděli od učitelů na SŠ, ze dne otevřených dveří a veletrhu Gaudeamus. Více než polovina studentů se rozhodla studovat pro pozdější vědeckou dráhu a vizi snadného získání VŠ diplomu.

Většina studentů ve všech shlucích se o možnosti studia dozvěděla od spolužáků a známých.

U Wardovy metody jsem zvolil hladinu se šesti shluky. První shluk obsahuje 369 objektů, druhý 247 objektů, třetí 256 objektů, čtvrtý shluk 382 objektů, pátý shluk 187 objektů a šestý shluk 223 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 61.

V prvním shluku jsou studenti, kteří byli v rozhodování ovlivněni exkurzí na pracovišti PřF UP a možností podílet se na výzkumu. O možnosti studia se studenti dozvěděli ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Více než polovina studentů se rozhodla studovat pro pozdější vědeckou dráhu.

Druhý shluk obsahuje studenty, které v rozhodování ovlivnil osobní kontakt se zaměstnancem nebo studentem PřF UP, možnost zůstat v místě bydliště, kamarádi, snadnost přijetí na PřF UP a nízké životní náklady v Olomouci. O možnosti studia se studenti dozvěděli z brožury PřF UP.

V třetím shluku se nachází studenti, kteří byli v rozhodování ovlivněni exkurzí na pracovišti PřF UP, možností zůstat v místě bydliště a podílet se na výzkumu. O možnosti studia se studenti dozvěděli ze dne otevřených dveří, brožury PřF UP. Naopak se většina nedozvěděla o možnosti studia z veletrhu Gaudeamus.

Čtvrtý shluk obsahuje studenty, kteří byli ovlivněni v rozhodování možností podílet se na výzkumu.

Pátý shluk obsahuje studenty, kteří byli v rozhodování ovlivněni exkurzí na pracovišti PřF UP, možností zůstat v místě bydliště, kamarády, snadností přijetí na PřF UP, možností podílet se na výzkumu a nízkými životními náklady v Olomouci. O možnosti studia se studenti dozvěděli od učitelů na SŠ, ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Více než polovina studentů se rozhodla studovat pro pozdější vědeckou dráhu.

V šestém shluku se nachází studenti, kteří nebyli v rozhodování většinou ovliv-

nění propagačními materiály. Část studentů také nebyla ovlivněna atraktivností příslušného oboru. Naopak byli ovlivněni snadností přijetí na PřF UP.

Většina studentů ve všech shlucích kromě prvního se o možnosti studia dozvěděla od spolužáků a známých.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 39. a 40. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	22 %	62 %	68 %
3	27 %	64 %	68 %
4	42 %	71 %	68 %
5	74 %	77 %	80 %
6	75 %	78 %	87 %

Tabulka 39.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	10 %	17 %	15 %
3	26 %	37 %	28 %
4	38 %	60 %	54 %
5	54 %	75 %	72 %
6	72 %	75 %	76 %

Tabulka 40.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

Procentuální podíly jsou u obou metod na přibližně stejné úrovni jako u shlukování na jednotlivých letech.

Data s obory

Stejně jako na datech z roku 2011 jsem připravil variantu, kdy jsem do dat pro shlukování přidal škálováním i informaci o oboru studenta.

Na tato data jsem opět aplikoval čtyři metody shlukování. Stejně jako v předchozí úloze shlukování vytvořila metoda complete linkage viz 27. a Wardova

metoda viz 28. vhodné dendrogramy pro další zkoumání. Ostatní dvě metody vytvořily dendrogramy obsahující v každé hladině dva shluky.

Typičtí zástupci jsou vybráni z datové matice, kde byly odstraněny ty řádky, kde počet neznámých hodnot odpovědí byl větší než 16. Při neodstranění těchto řádků se mezi typickými zástupci objevovali zástupci s příliš mnoha neznámými hodnotami.

U metody complete linkage jsem zvolil hladinu se čtyřmi shluky. První shluk obsahuje 226 objektů, druhý 762 objektů, třetí shluk 656 objektů a čtvrtý shluk 57 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 62.

V prvním shluku se nachází studenti, kteří byli v rozhodování ovlivněni možnostmi zůstat v místě bydliště, kamarády, snadností přijetí na PřF UP a nízkými životními náklady v Olomouci. O možnosti studia se dozvěděli ze dne otevřených dveří a brožury PřF UP. Studenti se také rozhodli studovat pro pozdější vědeckou dráhu.

Druhý shluk obsahuje studenty, které v rozhodování ovlivnily propagační materiály, možnost výzkumu a nízké životní náklady v Olomouci. O možnosti studia se dozvěděli ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP.

Ve třetím shluku se nachází skupina studentů, kteří se o možnosti studia nedozvěděli z internetových stránek PřF UP.

Čtvrtý shluk obsahuje studenty, které v rozhodování ovlivnili učitelé na SŠ, osobní kontakt se zaměstnancem nebo studentem PřF UP, propagační materiály, exkurze na pracovišti PřF UP, kamarádi, snadnost přijetí na PřF UP, možnost výzkumu, nízké životní náklady v Olomouci a vize nenáročného studia. O možnosti studia se dozvěděli od učitelů na SŠ, ze dne otevřených dveří, brožury PřF UP a z Facebooku či jiné sociální sítě. Studenti se rozhodli studovat pro pozdější vědeckou dráhu.

Většina studentů ve všech shlucích se o možnosti studia dozvěděla od spolužáků a známých.

U Wardovy metody jsem zvolil hladinu se šesti shluky. První shluk obsahuje 607 objektů, druhý 330 objektů, třetí 318 objektů, čtvrtý shluk 231 objektů, pátý shluk 107 objektů a šestý shluk 108 objektů. Typičtí zástupci shluků jsou uvedeni v tabulce 63.

První shluk obsahuje studenty, kteří byli při rozhodování ovlivněni exkurzí na pracovišti PřF UP a možnostmi výzkumu. O možnosti studia se dozvěděli ze dne otevřených dveří, veletrhu Gaudeamus a brožury PřF UP. Studenti se rozhodli studovat pro pozdější vědeckou dráhu.

Ve druhém shluku se nachází studenti, které při rozhodování ovlivnila možnost zůstat v místě bydliště, kamarádi, snadnost přijetí na PřF UP, možnost výzkumu a nízké životní náklady v Olomouci. V tomto shluku se nachází 78 studentů oboru Informatika z celkového počtu 97 studentů.

Studenti ve třetím shluku byli ovlivněni při rozhodování snadností přijetí na PřF UP.

Čtvrtý shluk obsahuje studenty, kteří byli při rozhodování ovlivněni exkurzí

na pracovišti PřF UP, snadností přijetí na PřF UP, možností výzkumu a nízkými životními náklady v Olomouci. O možnosti studia se studenti dozvěděli od učitelů na SŠ, ze dne otevřených dveří a brožury PřF UP.

V pátém shluku se nachází studenti, které při rozhodování ovlivnila možnost zůstat v místě bydliště, kamarádi, snadnost přijetí na PřF UP, možnost výzkumu a nízké životní náklady v Olomouci. V tomto shluku se nachází celkem 101 studentů z oboru Aplikovaná informatika z celkového počtu 124 studentů.

Studenti v šestém shluku se o možnosti studia dozvěděli od učitelů na SŠ, ze dne otevřených dveří a brožury PřF UP. Tento shluk tvoří 106 studentů z oboru Regionální geografie.

Většina studentů ve všech shlucích kromě třetího byla při rozhodování ovlivněna propagačními materiály. Většina studentů ve všech shlucích se o možnosti studia dozvěděla od spolužáků a známých.

Na data se změněnými vahami (váhy 0,25 a 0,75 a váhy 0,3 a 0,7) jsem opět aplikoval hierarchické shlukování. V tabulce 41. a 42. jsou uvedeny procentuální podíly jednotlivých vah u metody complete linkage a Wardovy metody. Původní hodnoty vah (hodnoty 0,2 a 0,8) odpovědí jsou označeny #1, váhy s hodnotami 0,25 a 0,75 jsou označeny #2 a váhy s hodnotami 0,3 a 0,7 jsou označeny #3.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	31 %	29 %	45 %
3	58 %	38 %	58 %
4	76 %	78 %	80 %
5	81 %	80 %	84 %
6	90 %	74 %	86 %

Tabulka 41.: Tabulka procentuálních podílů shluků jednotlivých vah u complete linkage.

Počet shluků	Podíl #1 a #2	Podíl #1 a #3	Podíl #2 a #3
2	16 %	35 %	38 %
3	46 %	51 %	22 %
4	47 %	51 %	22 %
5	58 %	56 %	22 %
6	66 %	51 %	56 %

Tabulka 42.: Tabulka procentuálních podílů shluků jednotlivých vah u Wardovy metody.

Procentuální podíly jsou u obou metod přibližně stejné popř. nižší než u shlukování bez zahrnutých oborů.

4.8.2. Metoda GUHA

Nastavení první úlohy je totožné jako nastavení předchozích prvních úloh. Jako kvantifikátor jsem nastavil BASE s absolutní hodnotou 10 a fundovanou implikaci s hodnotou 0,75. Takto zadaná úloha vygenerovala celkem 308 pravidel.

Výběr zajímavých pravidel:

- č. 1: ProcZajemObor(SpiseNe) >:< Kraj(Olomoucký)
- č. 2: ProcZajemObor(UrciteNe) >:< RozhodnutiRodice(UrciteNe)
- č. 3: RozhodnutiPropagacniMaterialy(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 4: RozhodnutiUcitele(UrciteAno) >:< ProcZajemObor(UrciteAno)
- č. 5: ZdrojeTelevizeNoviny(UrciteAno) >:< ZdrojeInternetPrf(UrciteAno)
- č. 6: StudijniObor(Bioanorganická chemie) >:< ZdrojeInternetPrf(UrciteAno)
- č. 7: StudijniObor(Biologie - zeměpis) >:< ProcDrahaVaV(SpiseNe)
- č. 8: StudijniObor(Ekochemie) >:< ZdrojeFacebook(UrciteNe)
- č. 9: StudijniObor(Optika a optoelektronika) >:< RozhodnutiNenarocneStudium(UrciteNe)
- č. 10: ZdrojeSoutezeSoustredeni(UrciteAno) >:< ProcZajemObor(UrciteAno)

Většina z uvedených pravidel se nachází v již uvedených výsledcích jednotlivých let. Pravidlo č. 8 říká, že studenti oboru Ekochemie se o možnosti studia nedozvěděli na Facebooku či jiné sociální síti. V budoucnu by tak bylo vhodné cílit na tyto studenty reklamními kampaněmi na sociálních sítích. Deváté pravidlo ukazuje, že studenti oboru Optika a optoelektronika si uvědomují, že studium jejich oboru bude náročné.

Pravidla č. 3, 4 a 5 se nachází také ve výsledcích dat z roku 2010. Dále všechna pravidla kromě č. 8 a 9 jsou obsažena ve výsledcích dat z roku 2011. Hodnoty podpory a spolehlivosti těchto pravidel jsou uvedeny v tabulce 43. Ve sloupci Podpora 2 a Spolehlivost 2 jsou uvedeny hodnoty pro data z roku 2010. Ve sloupci Podpora 3 a Spolehlivost 3 jsou uvedeny hodnoty pro data z roku 2011. Jak je z tabulky patrné, hodnoty podpory a spolehlivosti jsou velmi podobné u těchto společných pravidel. Pravidlo č. 8 uvádí, že studenti oboru Ekochemie se nedozvěděli o možnosti studia z Facebooku nebo jiné sociální sítě. Tento obor by tak mohl zveřejňovat reklamu na sociálních sítích, aby se studenti o možnosti studia na tomto oboru dozvěděli.

Hodnoty podpory a spolehlivosti pravidel jsou uvedeny v tabulce 43. Všechna uvedená pravidla kromě třetího dosahují malé podpory, která odpovídá desítkám záznamů.

Č. prav.	Podp.	Spolehl.	Podp. 2	Spolehl. 2	Podp. 3	Spolehl. 3
1	0,02	0,82	-	-	0,03	0,86
2	0,01	0,83	-	-	0,01	0,87
3	0,12	0,8	0,1	0,77	0,13	0,82
4	0,03	0,85	0,03	0,83	0,03	0,85
5	0,02	0,78	0,02	0,85	0,02	0,75
6	0,01	0,79	-	-	0,01	0,79
7	0,01	0,78	-	-	0,01	0,78
8	0,01	0,92	-	-	-	-
9	0,01	0,77	-	-	-	-
10	0,01	0,8	-	-	0,01	0,76

Tabulka 43.: Tabulka hodnot podpory a spolehlivosti.

4.9. Vyhodnocení výsledků

Při srovnání výsledků shlukování mezi daty oborů Aplikovaná informatika a Informatika mezi lety 2010 a 2011 vykazují shluky u metody complete linkage jistou míru podobnosti. Počet shluků je stejný, a tak lze jednotlivé shluky jednoduše porovnat. U Wardovy metody je počet shluků v roce 2011 o jeden menší než v roce 2010 a na první pohled nejsou ve shlucích vidět stejné vlastnosti. Výsledné shluky u sloučených dat z let 2010 a 2011 nevykazují u žádné z použitých metod jednoznačné související rysy při srovnání se shluky z jednotlivých let 2010 a 2011.

Srovnání výsledků shlukování na datech všech oborů mezi jednotlivými lety 2010 a 2011 neprokazuje vysokou míru podobnosti shluků u žádné z provedených metod. U dat s obory je situace u metody complete linkage totožná. U Wardovy metody je druhý shluk dle charakteristiky totožný mezi daty s obory. Shluk obsahuje v případě roku 2010 celkem 88 objektů a v roce 2011 celkem 399 objektů.

Mezi daty z roku 2011 a sloučenými daty roku 2010 a 2011 jsou u metody complete linkage znatelné společné vlastnosti mezi druhým a třetím shlukem a dále mezi třetím a druhým shlukem. U Wardovy metody nejsou společné znaky shluků viditelné.

U metody complete linkage a Wardovy metody mezi daty z roku 2010 a sloučenými daty z roku 2010 a 2011 jsou podobné pouze první shluky. Ostatní shluky mají mezi sebou různé vlastnosti.

Srovnání jednotlivých nalezených pravidel z metody GUHA je uvedeno pod jednotlivými výsledky metody v textu. V tomto srovnání lze vidět rozdíly mezi hodnotami podpory a spolehlivosti na jednotlivých datech. Data z roku 2011 obsahují téměř dvakrát více dat (počet objektů) z dotazníků než z roku 2010. Díky tomuto faktu mají výsledky z roku 2011 a výsledky ze sloučených dat 2010

a 2011 větší počet společných pravidel.

4.9.1. Interpretace pro zákazníka

Interpretace výsledků z hierarchického shlukování může být zákazníkovi prezentována pomocí charakteristiky jednotlivých shluků z dat roku 2010, 2011, dále sloučených dat roku 2010 a 2011 a také pouze z dat oborů Aplikovaná informatika a Informatika. Pro shlukování byly významné dvě metody - complete linkage a Wardova metoda. Pro srozumitelnost interpretace bych preferoval prezentaci pouze výsledků Wardovy metody, protože výsledné shluky byly na dendrogramu dobře formované a charakteristika shluků byla provedena na takovém počtu shluků, kde byla hladina rozdílů největší. Pro detailnější charakterizaci výsledků by bylo možné použít i hladiny s menšími rozdíly, zvláště u dat, kde byla do shlukování zahrnuta i informace o oboru studenta.

Výsledky prezentace charakteristik shluků by mohl zákazník využít pro další účely, jako např. cílení reklamních kampaní na Facebooku jen pro určité obory, zvýšení distribuce propagačních materiálů v jiných krajích nebo větší informovanost o konání dne otevřených dveří či veletrhu Gaudeamus.

Výsledky z metody GUHA jsou pro zákazníka ve formě asociačních pravidel snadno srozumitelné. Doplnující interpretace jsou uvedeny v textu u každých výsledků metody GUHA. Nalezená pravidla nemají obecně velkou podporu, a tak platí pro relativně malou část ze všech studentů. Níže jsou uvedena vybraná nejzajímavější pravidla z výsledků metody GUHA na datech oborů Aplikovaná informatika a Informatika ze všech provedených pokusů:

- č. 1: ProcSnadnyDiplom(SpiseAno) >:< ZdrojeSpoluzaciZnami(SpiseAno)
- č. 2: RozhodnutiRodice(SpiseAno) >:< RozhodnutiVyzkum(SpiseAno)
- č. 3: ZdrojeFacebook(SpiseAno) >:< RozhodnutiPrestizPrf(SpiseAno)
- č. 4: ZdrojeFacebook(SpiseAno) >:< RozhodnutiVyzkum(SpiseAno)

Výběr nejzajímavějších pravidel z dat všech oborů ze všech provedených pokusů jsou uvedeny níže:

- č. 1: ProcSnadnyDiplom(UrciteAno) >:< Kraj(Olomoucký)
- č. 2: StudijniObor(Optometrie) >:< ProcPrace(UrciteAno)
- č. 3: ZdrojePrezentaceNaSs(UrciteAno) >:< ProcZajemObor(UrciteAno)
- č. 4: RozhodnutiSpolupraceProjekty(UrciteAno) >:< Kraj(Olomoucký)
- č. 5: Kraj(Vysočina) >:< RozhodnutiPrestizPrf(SpiseAno)
- č. 6: ProcZajemObor(SpiseNe) >:< Kraj(Olomoucký)
- č. 7: StudijniObor(Biologie - zeměpis) >:< ProcDrahaVaV(SpiseNe)
- č. 8: StudijniObor(Ekochemie) >:< ZdrojeFacebook(UrciteNe)

Hodnoty podpory a spolehlivosti těchto pravidel jsou uvedeny ve výsledcích metody GUHA, kde se tyto pravidla vyskytují. Výše uvedená pravidla považují za nová a dříve neznámá, a tak použití metody GUHA hodnotím kladně.

4.10. Závěr

Na data z dotazníků při zápisu do 1. ročníku Přírodovědecké fakulty UP v Olomouci v roce 2010 a 2011 jsem celkově aplikoval tři data miningové metody.

Použití asociačních pravidel nepovažuji za vhodné, jelikož nalezených pravidel je velké množství a neustále se opakují různé triviální kombinace. Při použití kvantifikátoru fundované implikace jsou asociační pravidla vytvářena i metodou GUHA.

Hierarchické shlukování vytváří ze své podstaty shluky objektů, které mají společné vlastnosti. Při aplikování na data z dotazníků se tak vytvořily shluky, které seskupují studenty dle jejich odpovědí a případně i jejich studijního oboru. V průběhu shlukování se však objevilo mnoho otázek ohledně správné volby vah odpovědí a metriky. V datech byly objeveny tzv. jádra shluků a fluktuanti. Jádra shluků jsou ty objekty, které nebyly volbou různých vah ovlivněny. Fluktuanti jsou naopak objekty, které se objevily při jiné váze v jiném shluku. Jedna z možných odpovědí je vysvětlení, že jádra shluků nemají tendenci volit příliš mnoho odpovědí spíše ano a spíše ne. Tato domněnka se však při pohledu na zdrojová data nepotvrdila.

Vysoké procentuální podíly mezi různými vahami odpovědí v dotaznících zpochybňují, ve velké míře, výsledky samotného shlukování. Volba vah a metriky se ukázala jako velice zásadní a důležité kritérium, které by bylo vhodné v dalším pokračování práce blíže prozkoumat.

Na data jsem aplikoval celkem čtyři metody shlukování. Metody average a single linkage nevytvořily v žádné úloze vhodné dendrogramy pro další zkoumání. Naproti tomu metody complete linkage a Wardova metoda vytvořily dendrogramy, které jsem dále analyzoval. Wardova metoda vytvářela vždy velmi dobře rozpoznatelné a oddělené shluky.

Při zkoumání shluků všech oborů u jednotlivých let jsem si všiml specifického umístění studentů stejného oboru. U shluků všech oborů na datech z let 2010 a 2011 se však toto umístění studentů nezachovalo. Větší váhu přikládám shlukům z těchto sloučených dat jednotlivých let, protože obsahují větší množství záznamů.

Využití hierarchického shlukování na dotaznících vidím ve schopnosti rozdělit studenty do skupin (shluků) a podle jejich charakteristických vlastností např. určit marketingové strategie pro jednotlivé obory. Z dendrogramů jsem vždy vybral takové shluky, které měly největší rozdíly mezi sousedními hladinami. Pro detailnější zkoumání by bylo možné vybrat větší počet shluků.

Metoda GUHA je velmi propracovaná a spolu se softwarem LISp-Miner tvoří silný nástroj pro generování hypotéz. V datech se podařilo najít množství pravidel, z nichž jsem vybral ty, které se mi zdály jako zajímavé. Obecně ale měla tato vybraná pravidla nízkou hodnotu podpory, a proto platí pro poměrně malou část studentů.

Závěr

Problematika shlukování je velmi rozsáhlá. V této diplomové práci jsem využil asociační pravidla, hierarchické shlukování a metodu GUHA na dotaznících při zápisu do 1. ročníku na Přírodovědeckou fakultu UP v Olomouci. U hierarchického shlukování jsem analyzoval jednotlivé shluky. Pomocí metody GUHA jsem vygeneroval a vybral zajímavá pravidla o studentech. Analyzované shluky i pravidla se dají použít pro další účely (např. marketingové).

Hlavním přínosem práce pro mě bylo hlubší poznání problematiky dolování dat a ověření smysluplnosti metodiky CRISP-DM. V případě dalšího výzkumu v této oblasti by bylo vhodné se hlouběji zaměřit na problematiku vah odpovědí a metriky. Vzhledem k velkému množství metod nehierarchického shlukování by stálo za pokus aplikovat tyto metody na data z dotazníků.

Conclusions

The clustering is a very extensive theme. In this master thesis I would like to apply associative rules, hierarchic clustering and the GUHA method to questionnaires from the students who subscribed for the first year of study on the Faculty of Science, Palacky University in Olomouc. I have analyzed separate clusters in hierarchic clustering. By using the GUHA method I have generated and chosen interesting rules about the students. Analyzed clusters and rules can be used for other purposes, e.g. marketing.

The main benefit of the thesis for myself was deeper understanding of the problems of quarrying data and validation of CRISP-DM method. For the purpose of further research, it would be appropriate to focus on the weight of answers and metrics. Considering the big amount of non-hierarchic clustering methods it might be useful to apply these methods on the data from questionnaires.

Reference

- [1] Jana Šarmanová. *Metody dolování znalostí z dat*. Datakon Brno, 2002.
- [2] Jana Šarmanová. *Metody analýzy dat*. [učební text stejnojmenného předmětu], VŠB-TU, 2002.
- [3] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth. *CRISP-DM 1.0 – Step-by-step data mining guide*. Dostupné z: <ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf>, staženo 7. 12. 2011
- [4] Petr Berka. *Dobývání znalostí z databází*. Academia, 2003.
- [5] Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.
- [6] Rakesh Agrawal, Ramakrishnan Srikant. *Fast Algorithms for Mining Association Rules*. Dostupné z: <http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf>, staženo 28. 5. 2012
- [7] Michal Burda. *Získávání znalostí z databází - Asociační pravidla*. Dostupné z: <http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0304/GUHA-text.pdf>, staženo 13. 2. 2012
- [8] Petr Hájek, Tomáš Havránek, Metoděk K. Chytil. *Metoda GUHA - automatická tvorba hypotéz*. Academia Praha, 1983.
- [9] Leonard Kaufman, Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 2005.
- [10] Sergios Theodoridis, Konstantinos Koutroumbas. *Pattern Recognition*. Academic Press, 2008.
- [11] Marija Norusis. *SPSS 17.0 Statistical Procedures Companion*. Prentice Hall, 2010.
- [12] Milan Šimůnek. *Systém LISp-Miner: akademický systém pro dobývání znalostí z databází*. Oeconomica, 2010
- [13] Jan Rauch. *Systém LISp-Miner*. Dostupné z: http://www.pejcoch.com/vyuka/4IZ210/ukol3/LM-SKRPT_11.pdf, staženo 5. 6. 2012
- [14] R. <http://www.r-project.org/>
- [15] William N. Venables, Brian D. Ripley. *Modern Applied Statistics with S*. Springer, 2002.

A. Obsah přiloženého CD

doc/

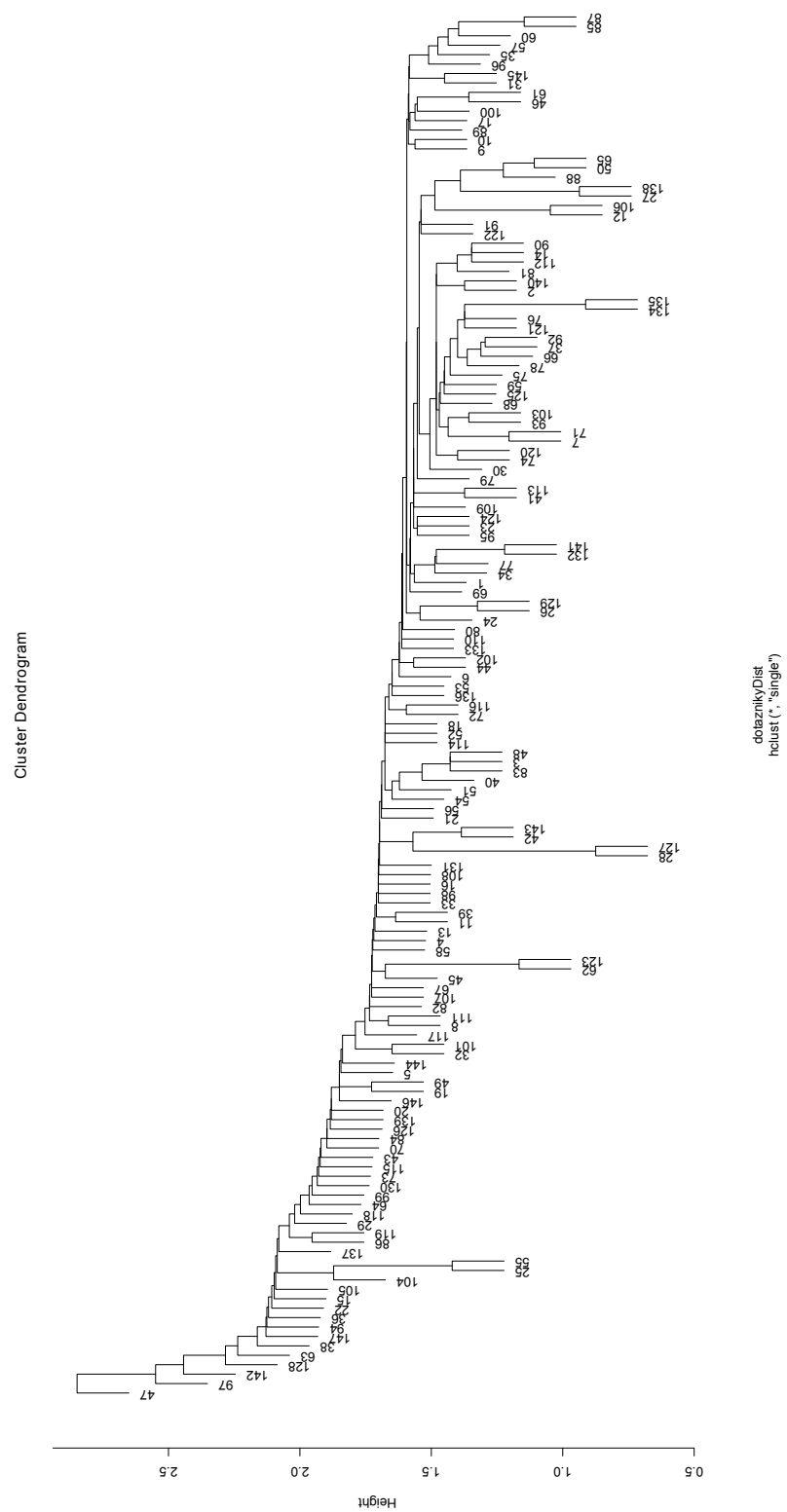
Dokumentace práce ve formátu PDF, vytvořená dle závazného stylu KI PřF pro diplomové práce, včetně všech příloh, a všechny soubory nutné pro bezproblémové vygenerování PDF souboru dokumentace, tj. zdrojový text dokumentace, vložené obrázky, apod.

dotazniky/

Adresář obsahující vstupní data a zdrojové soubory dolování dat dotazníků při zápisu do 1. ročníku PřF UP v Olomouci.

U veškerých odjinud převzatých materiálů obsažených na CD jejich zahrnutí dovolují podmínky pro jejich šíření.

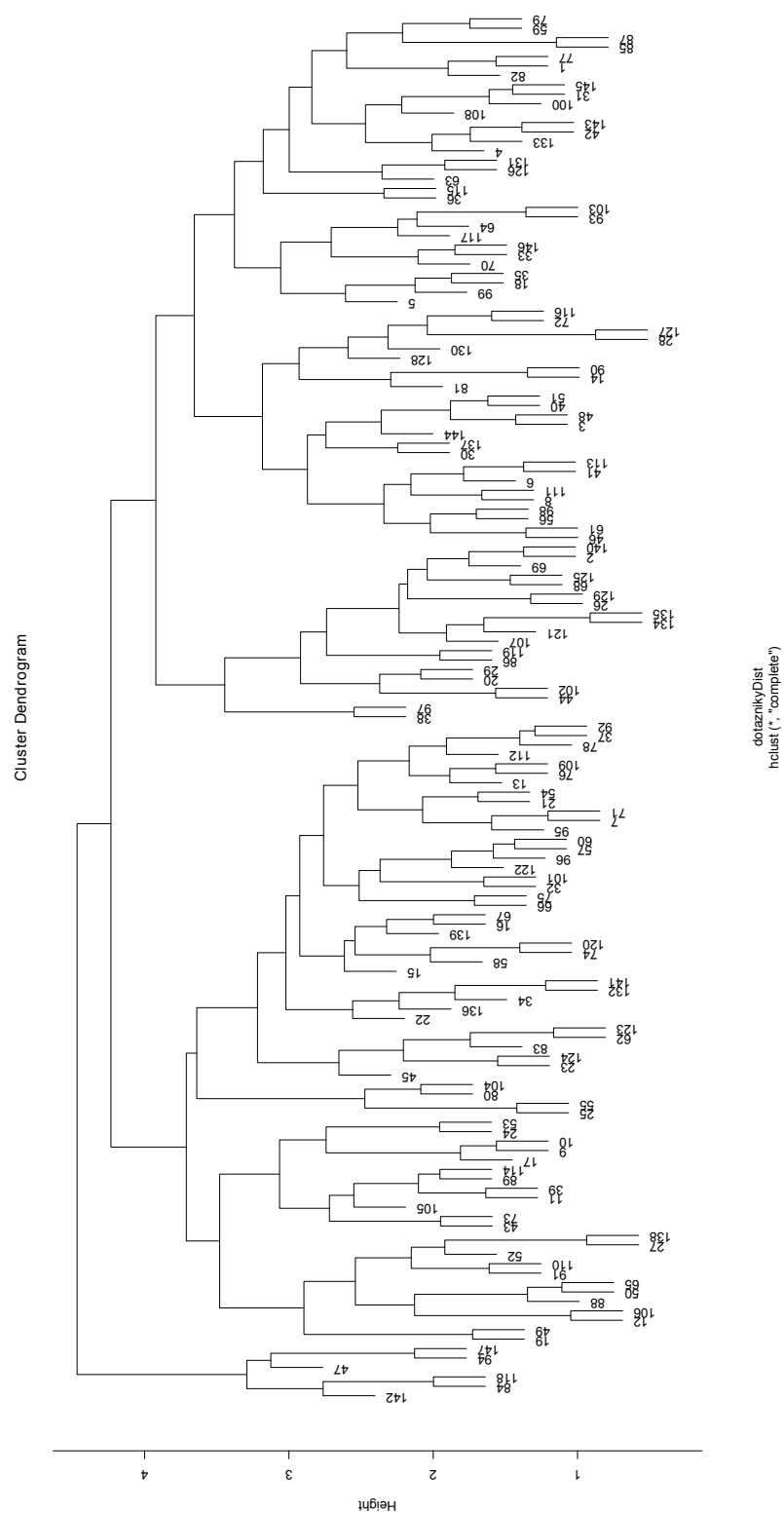
B. Dendrogramy hierarchického shlukování



Obrázek 7.: Dendrogram shlukování - Apl. inf. a Informatika 2011 (single linkage)



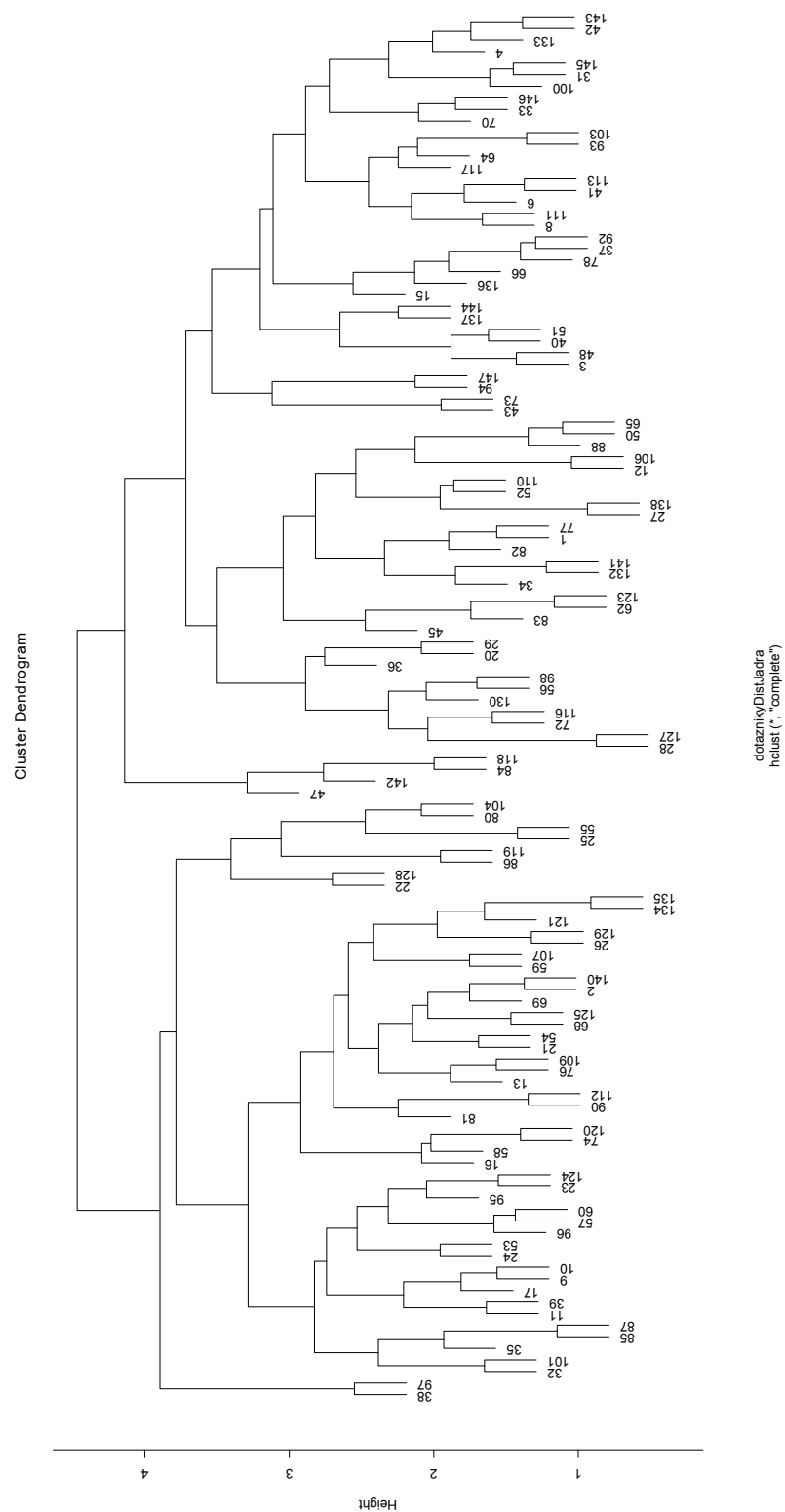
Obrázek 8.: Dendrogram shlukování - Apl. inf. a Inf. 2011 (average linkage)



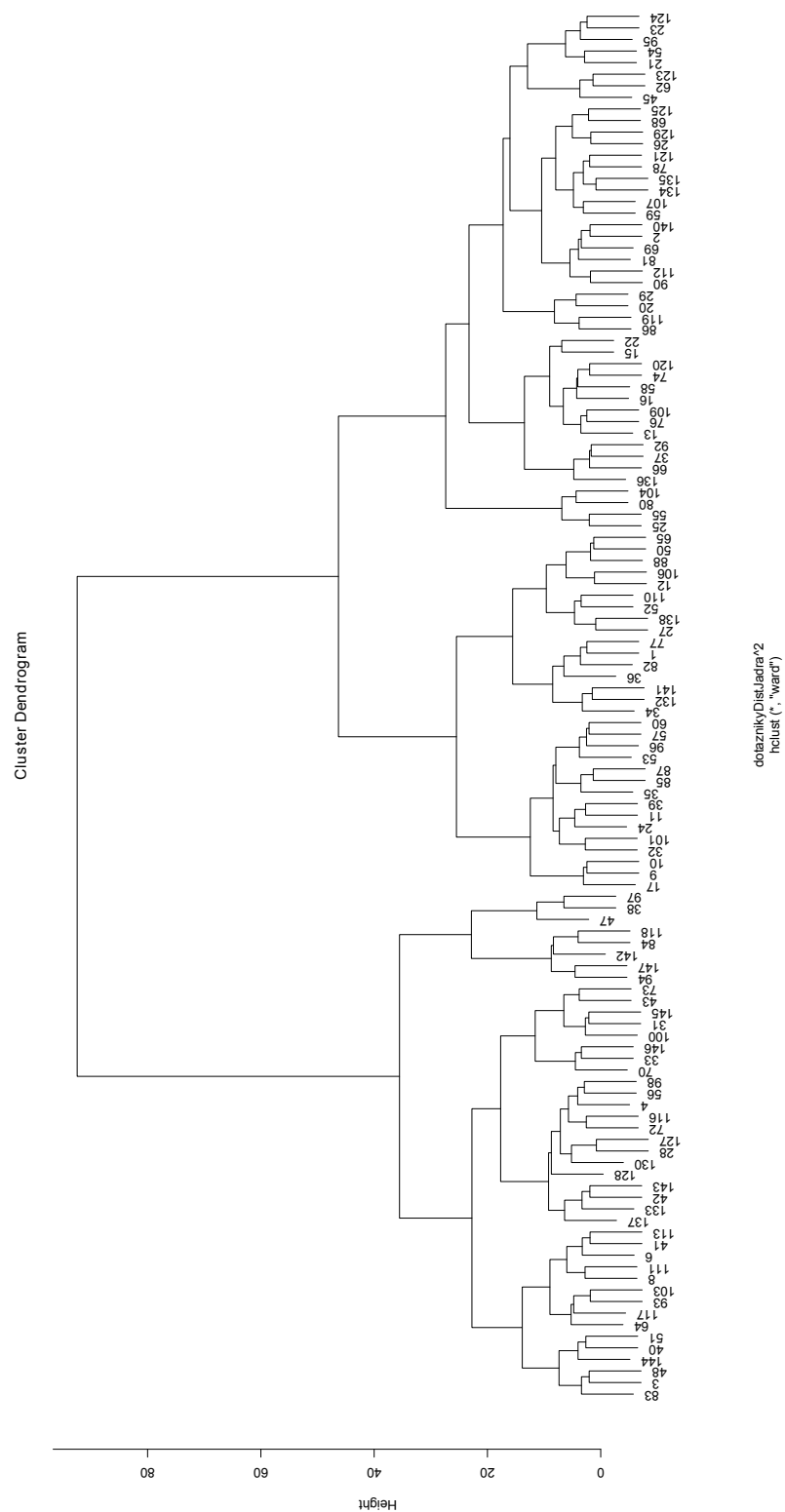
Obrázek 9.: Dendrogram shlukování - Apl. inf. a Inf. 2011 (complete linkage)



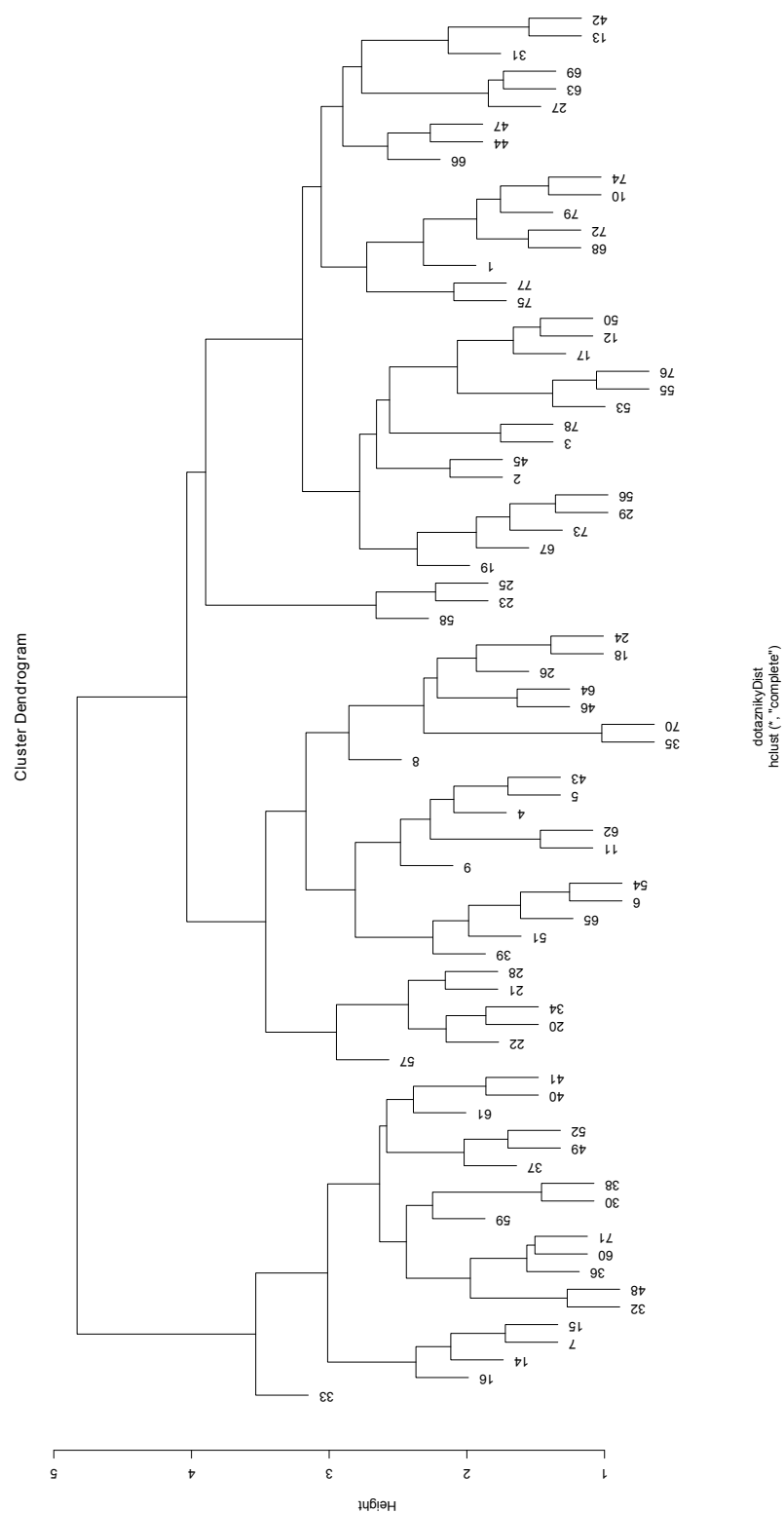
Obrázek 10.: Dendrogram shlukování - Apl. inf. a Inf. 2011 (Wardova metoda)



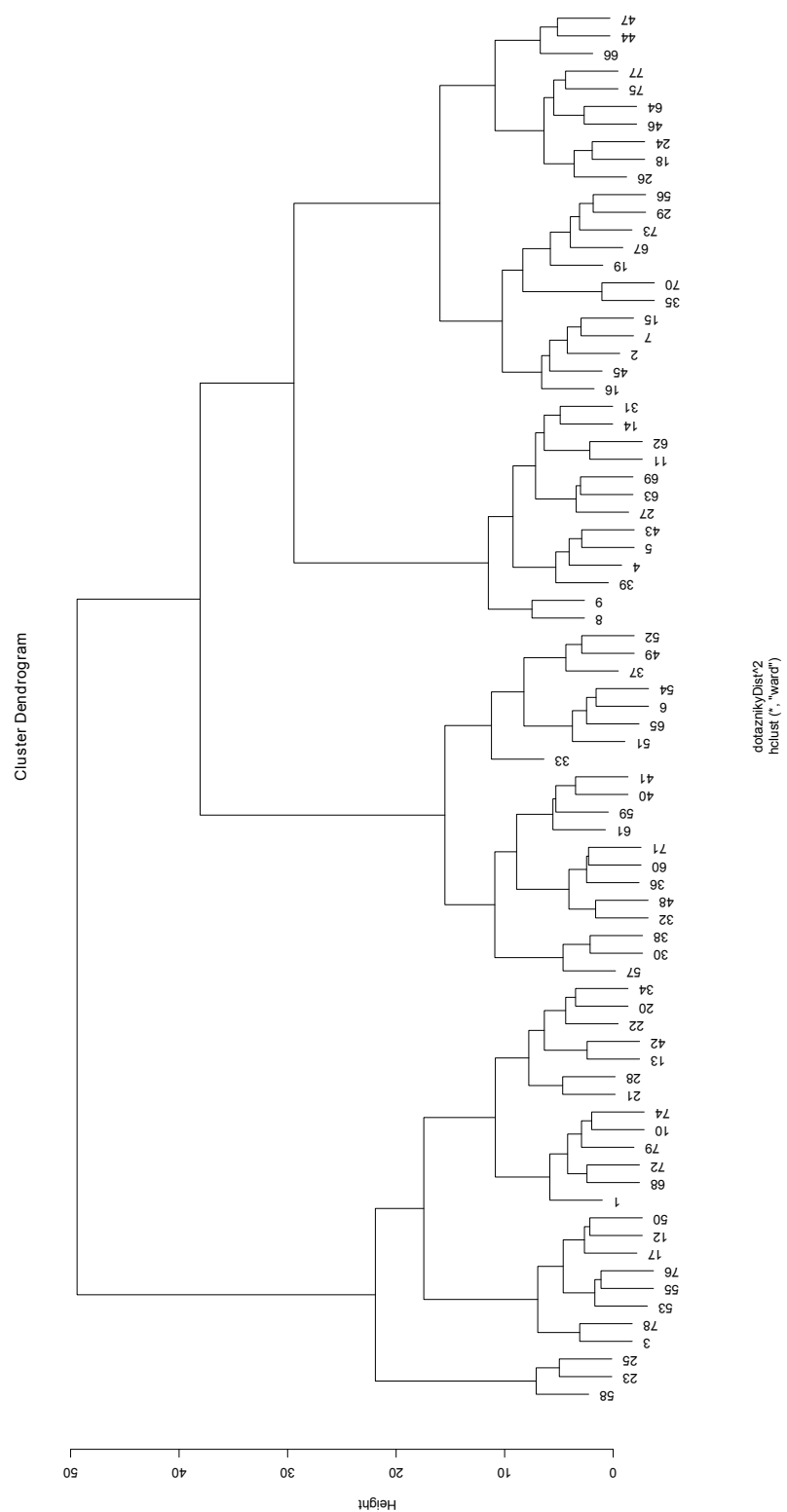
Obrázek 11.: Dendrogram shluk. - Apl. inf. a Inf. 2011 - jádra (complete linkage)



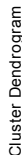
Obrázek 12.: Dendrogram shluk. - Apl. inf. a Inf. 2011 - jádra (Wardova metoda)



Obrázek 13.: Dendrogram shlukování - Apl. inf. a Inf. 2010 (complete linkage)



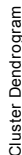
Obrázek 14.: Dendrogram shlukování - Apl. inf. a Inf. 2010 (Wardova metoda)

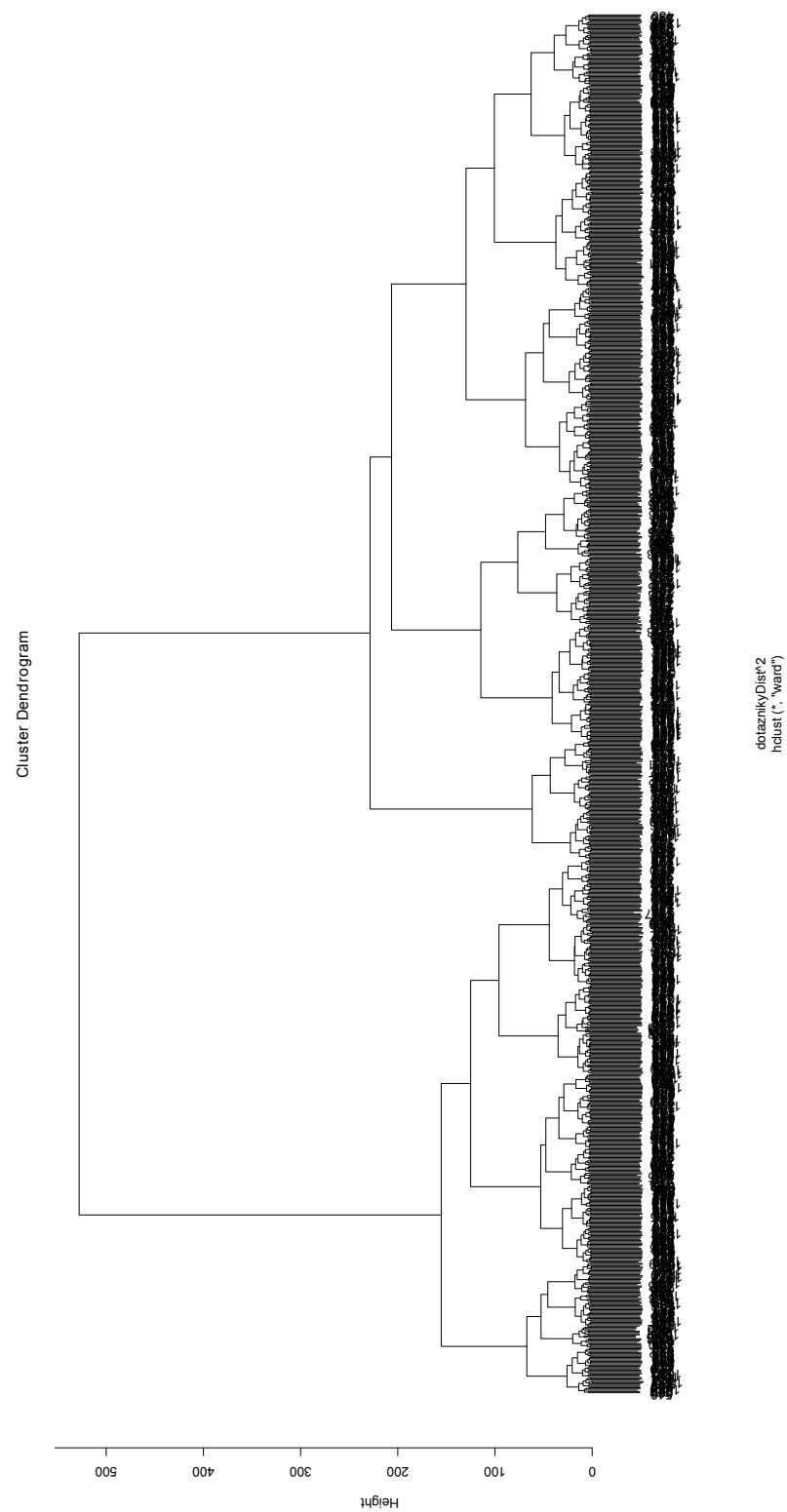


83

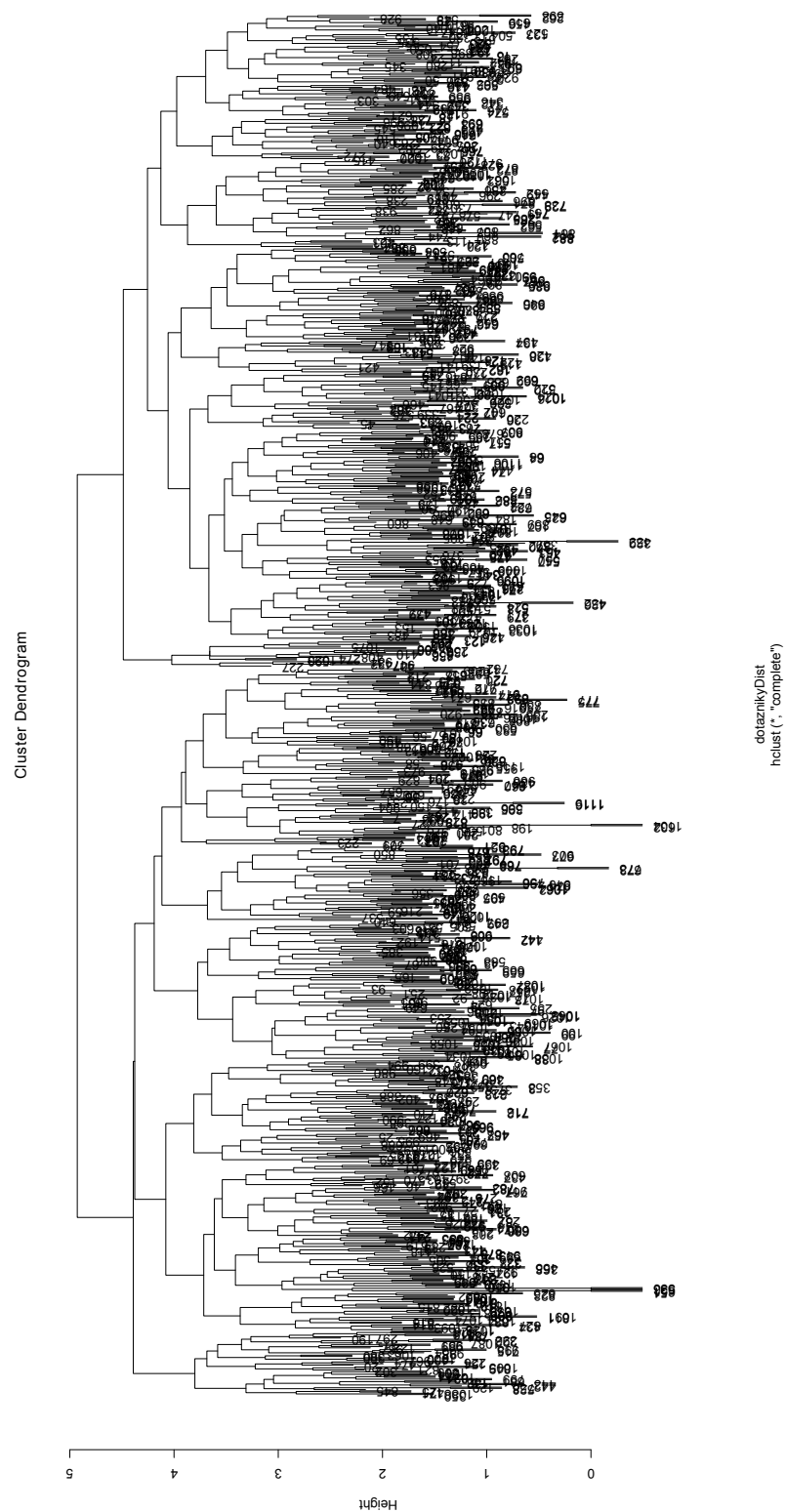


Obrázek 16.: Dendrogram shluk. - Apl. inf. a Inf. 2010 a 2011 (Wardova metoda)

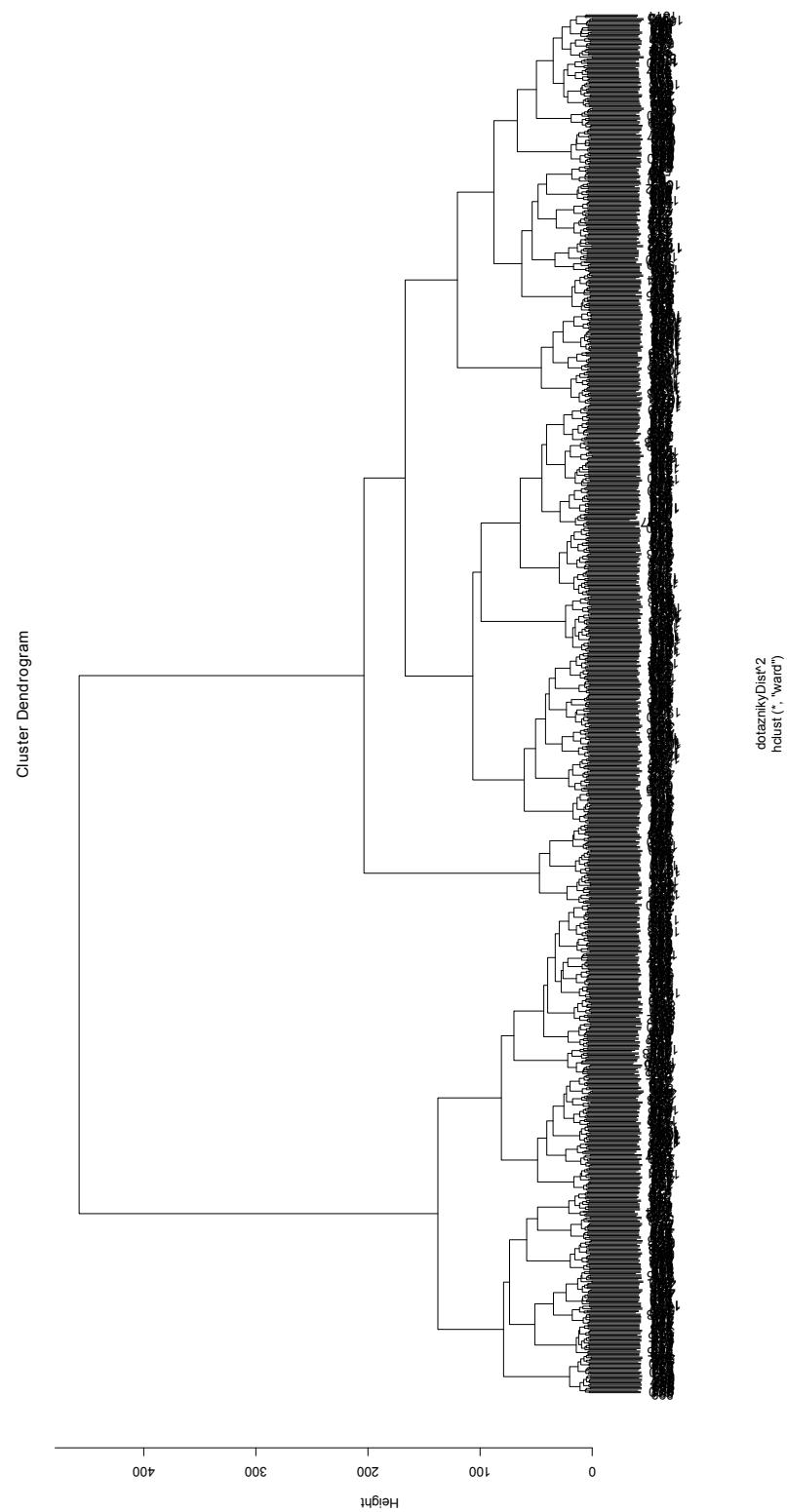
dotaznikyDist
includ (*, "complete")



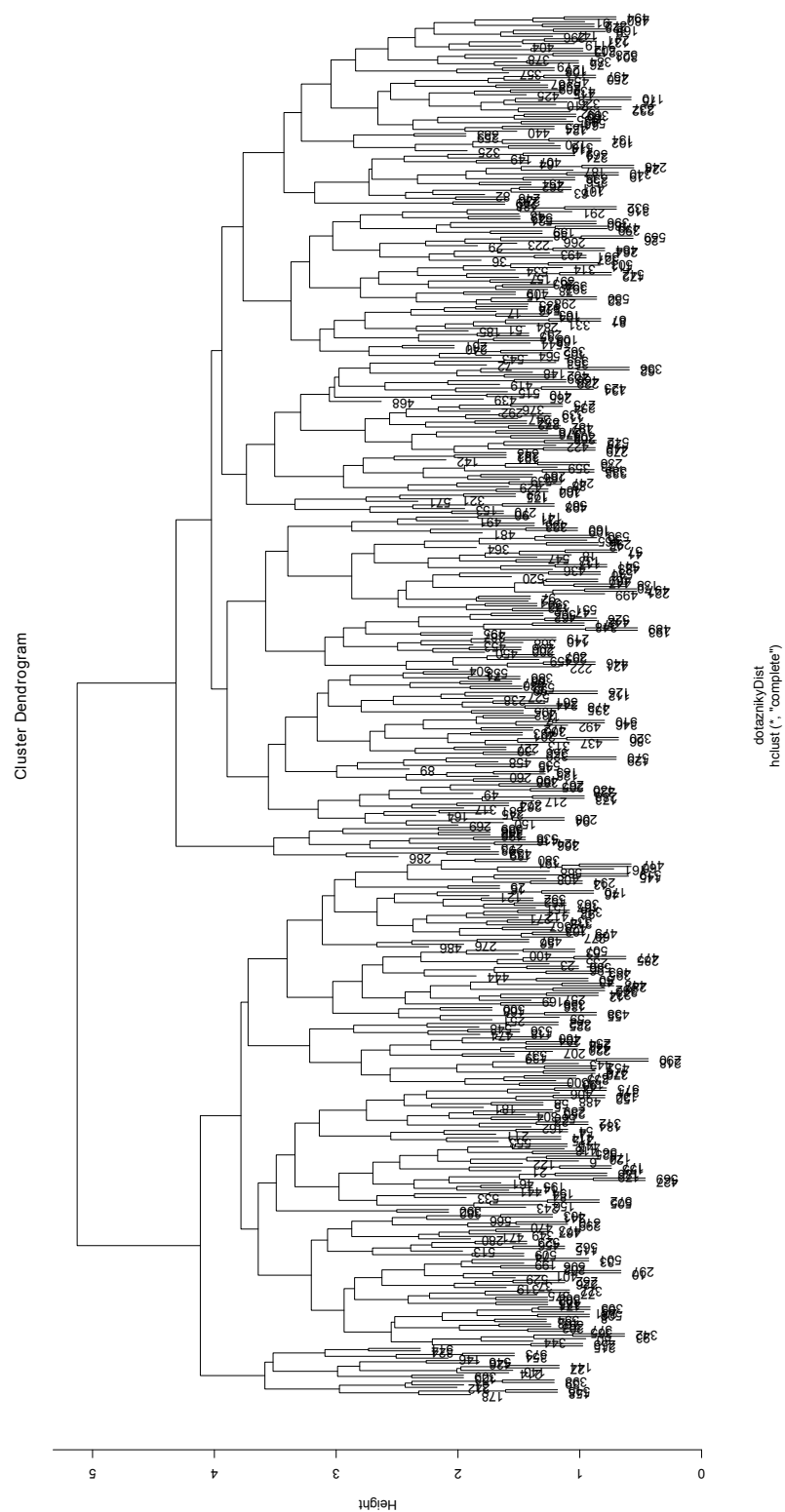
Obrázek 18.: Dendrogram shlukování - 2011 (Wardova metoda)



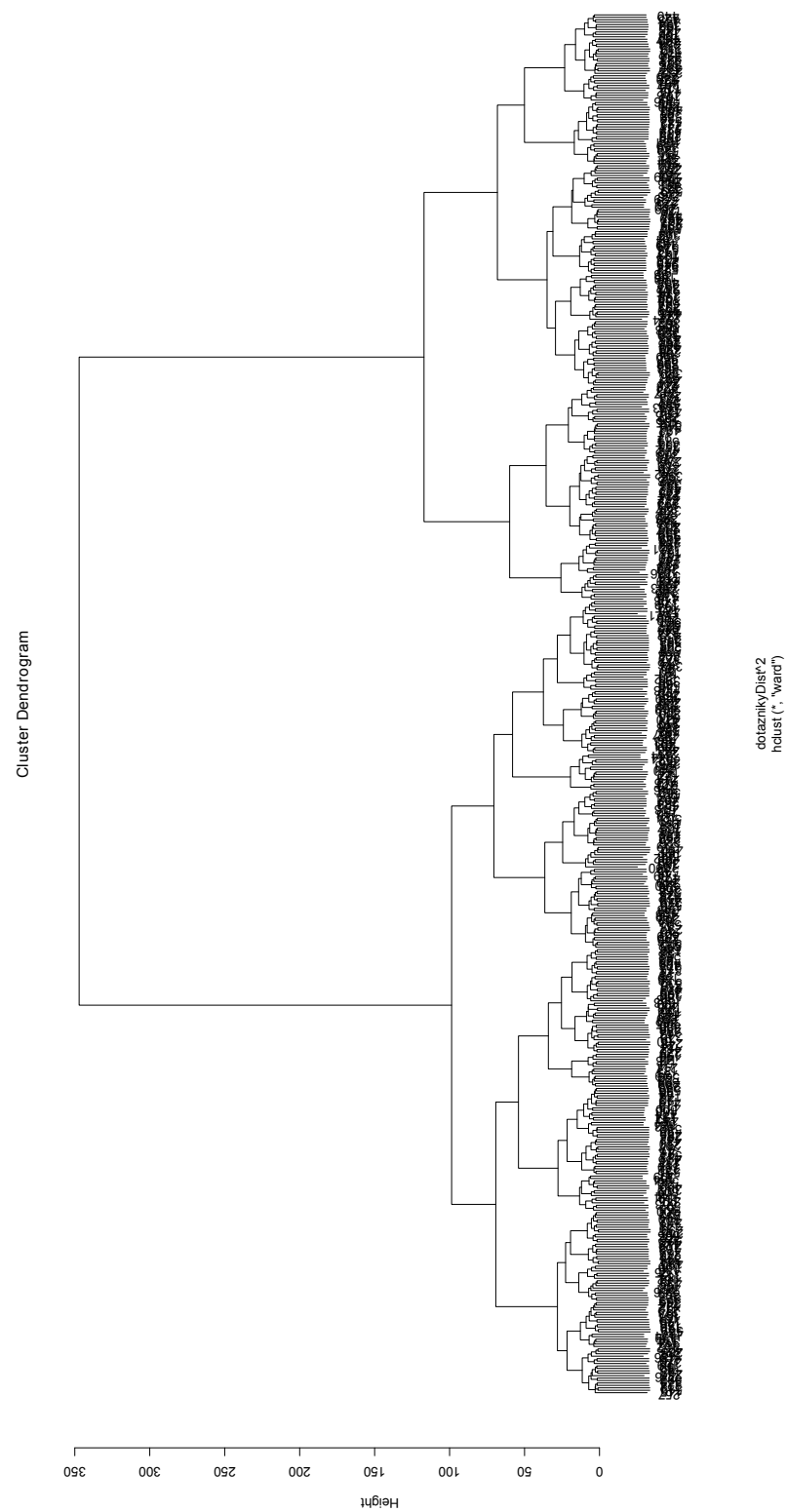
Obrázek 19.: Dendrogram shluk. s obory - 2011 (complete linkage)



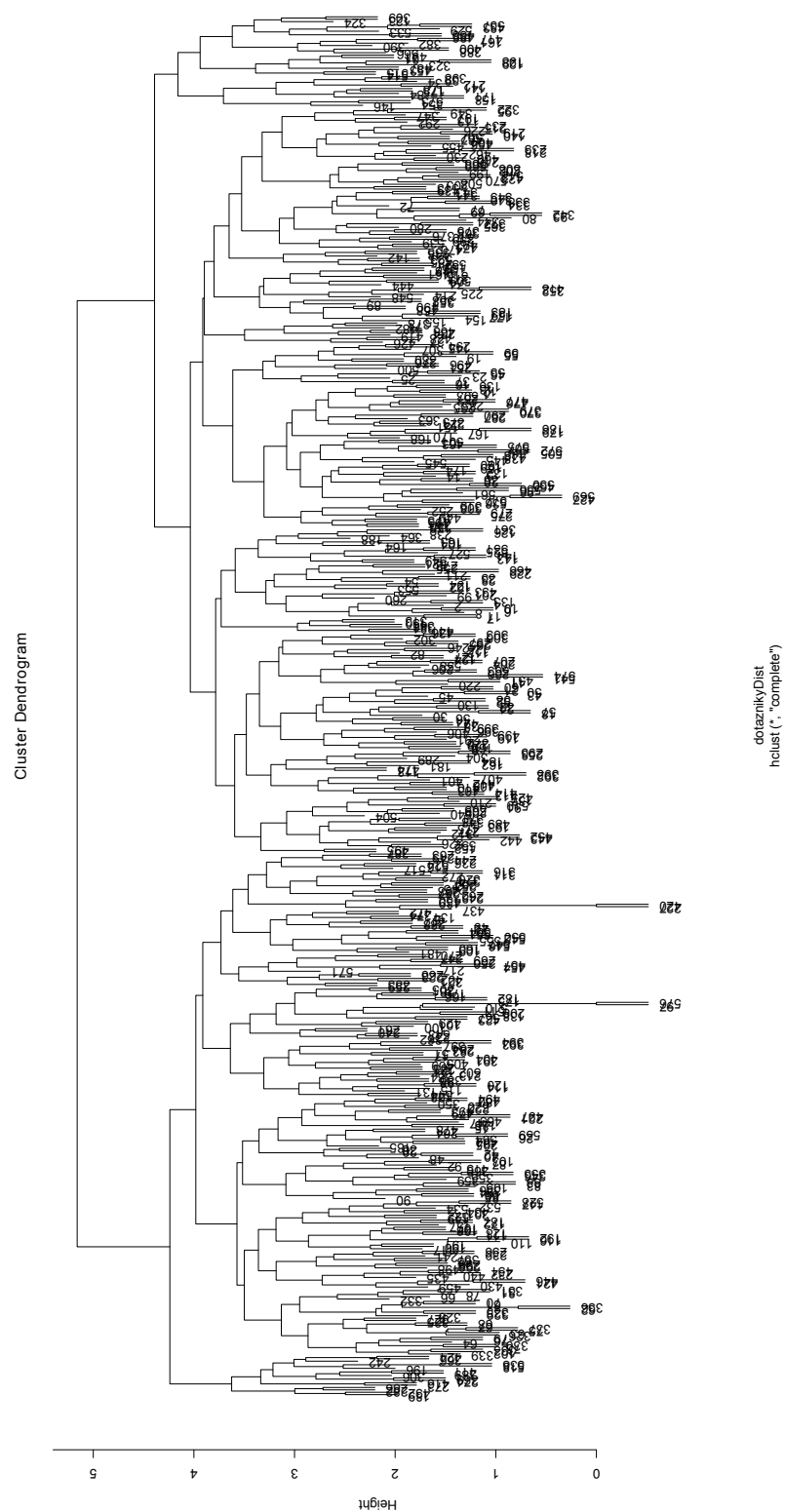
Obrázek 20.: Dendrogram shluk. s obory - 2011 (Wardova metoda)



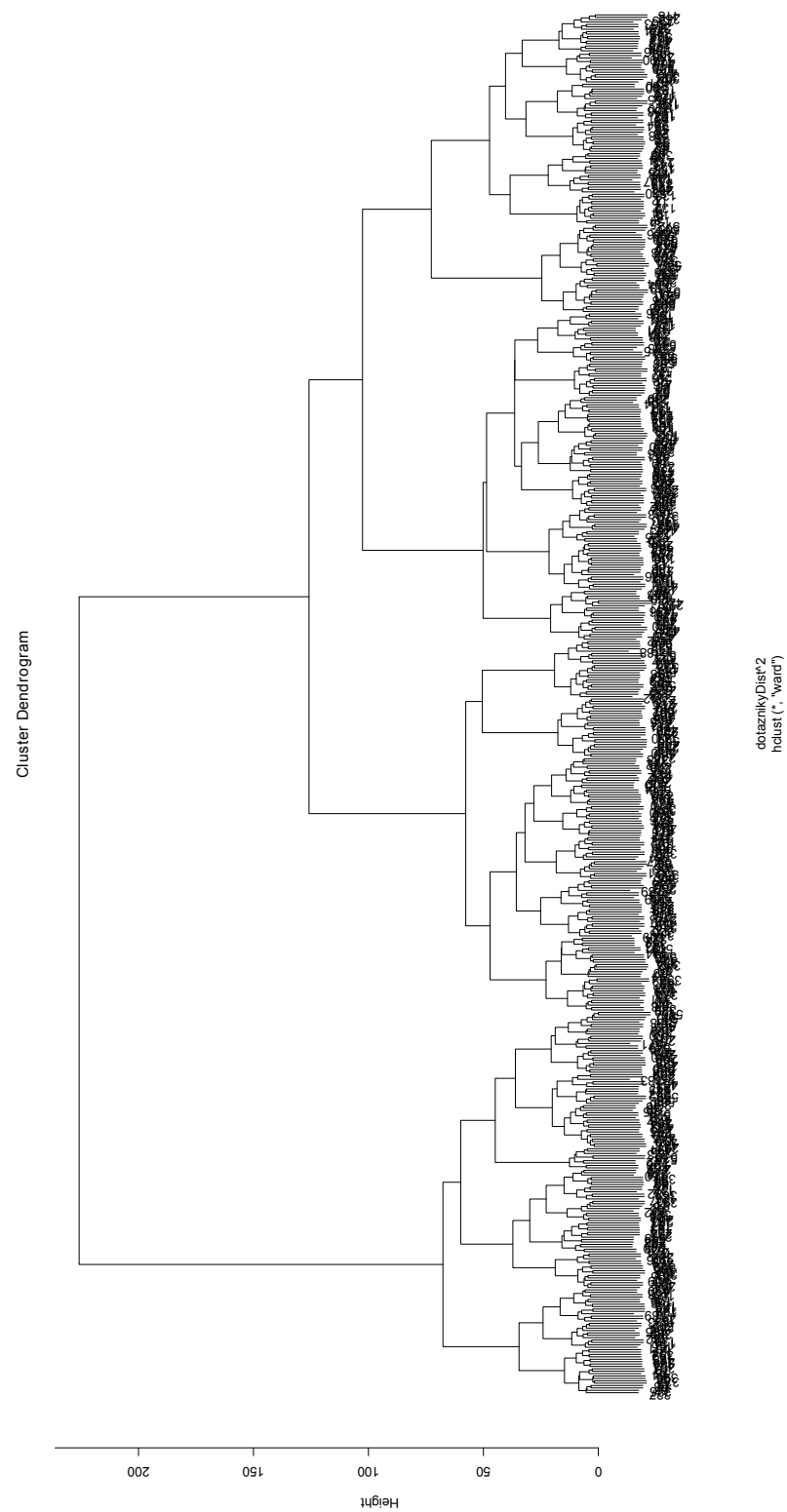
Obrázek 21.: Dendrogram shlukování - 2010 (complete linkage)



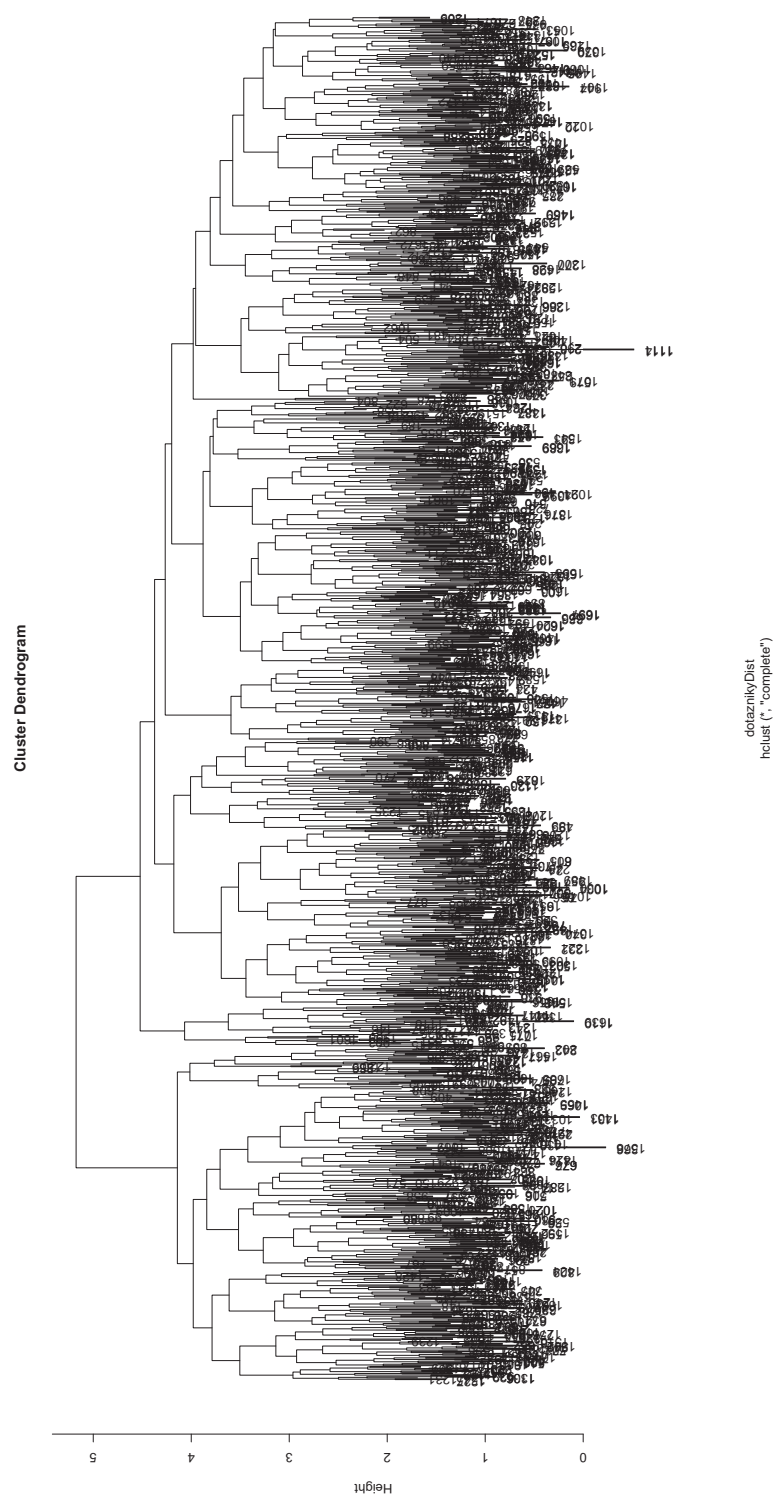
Obrázek 22.: Dendrogram shlukování - 2010 (Wardova metoda)



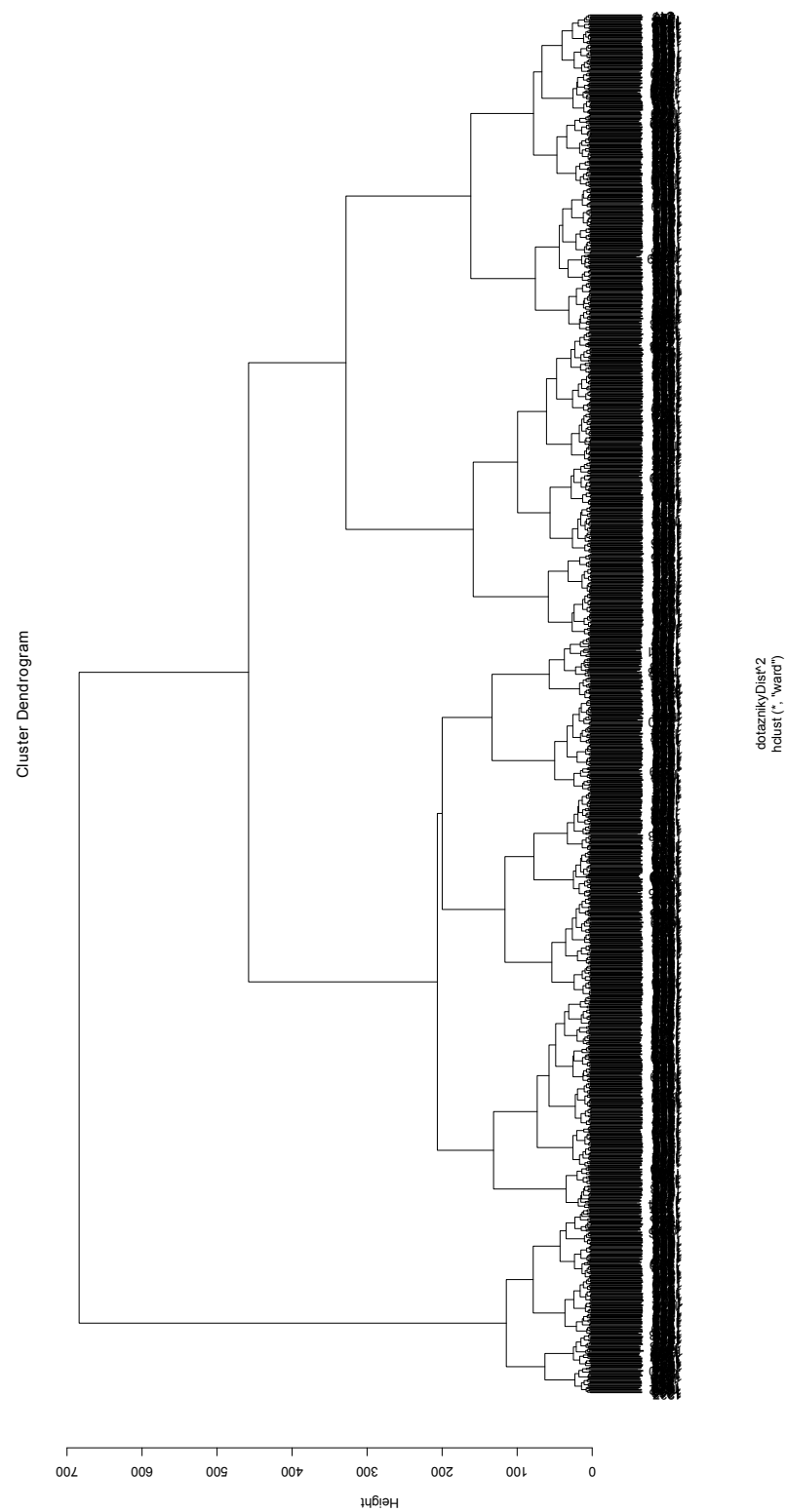
Obrázek 23.: Dendrogram shluk. s obory - 2010 (complete linkage)



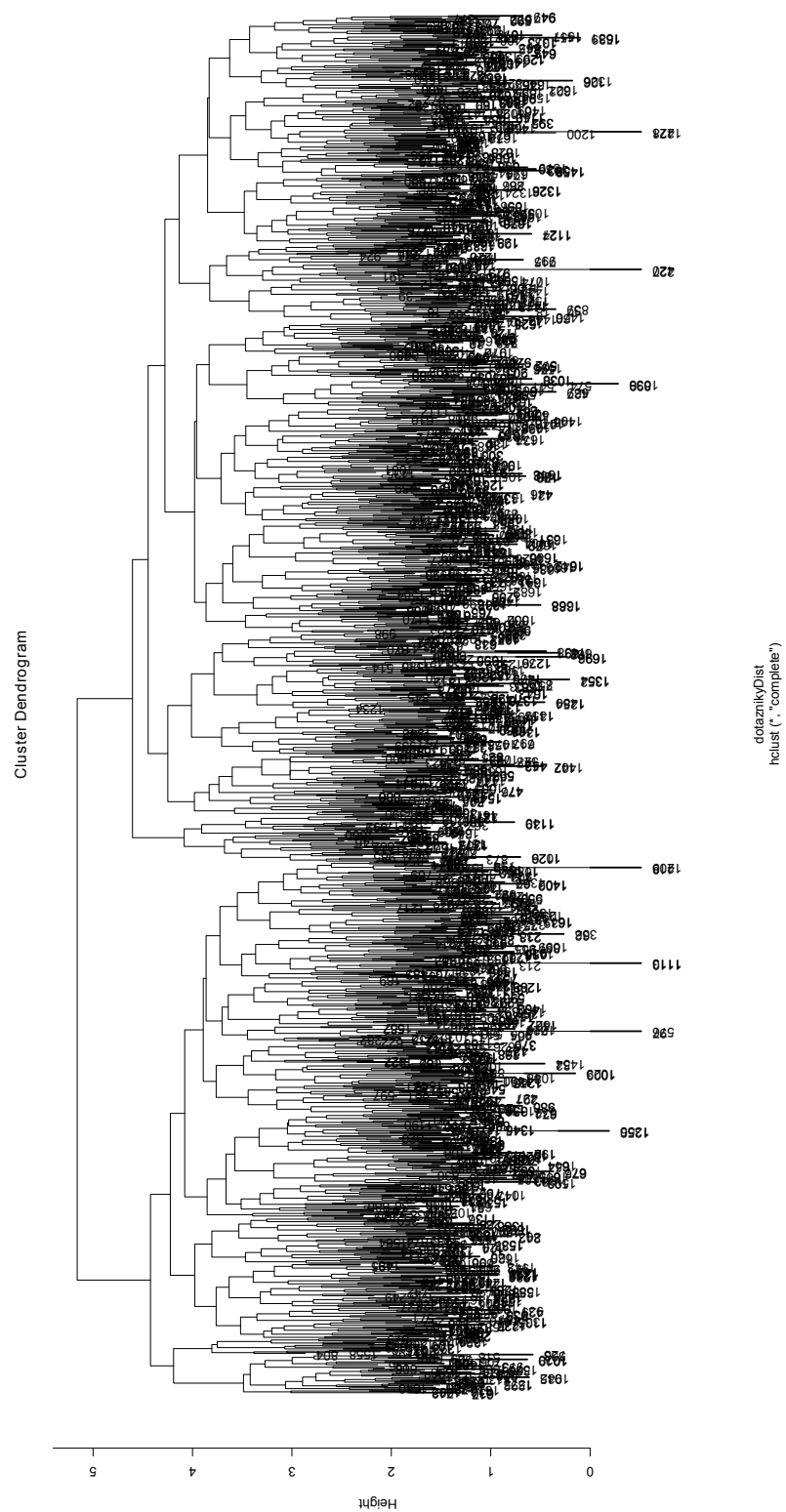
Obrázek 24.: Dendrogram shluk. s obory - 2010 (Wardova metoda)



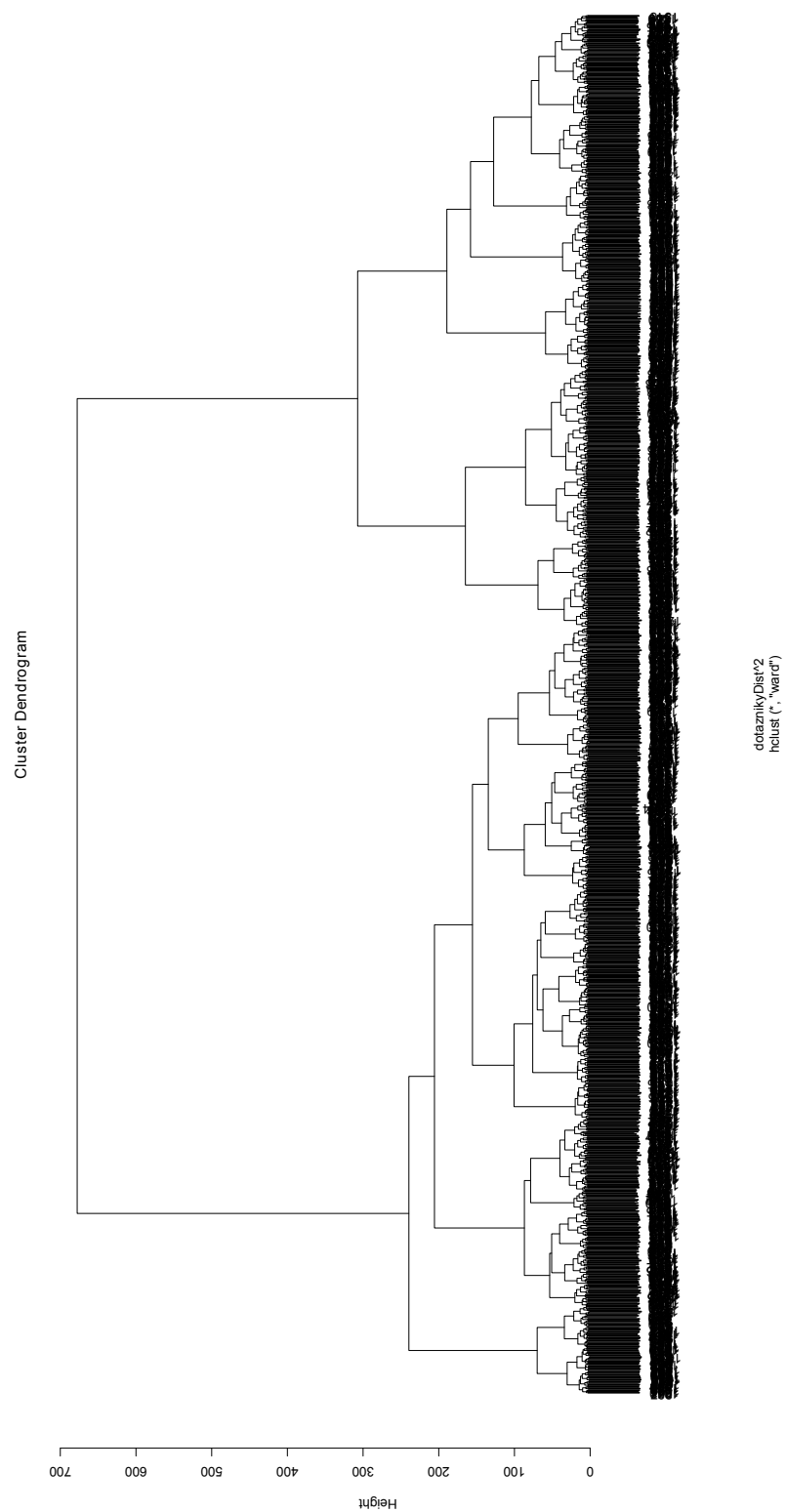
Obrázek 25.: Dendrogram shlukování - 2010 a 2011 (complete linkage)



Obrázek 26.: Dendrogram shlukování - 2010 a 2011 (Wardova metoda)



Obrázek 27.: Dendrogram shluk. s obory - 2010 a 2011 (complete linkage)



Obrázek 28.: Dendrogram shluk. s obory - 2010 a 2011 (Wardova metoda)

C. Typičtí zástupci shluků

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Spíše ano	Spíše ne	Spíše ne
2	Spíše ne	Určitě ne	Určitě ne	Spíše ne
3	Spíše ano	Spíše ne	Spíše ne	Spíše ano
4	Spíše ne	Určitě ne	Spíše ne	Spíše ne
5	Spíše ano	Spíše ano	Spíše ne	Určitě ano
6	Spíše ano	Určitě ne	Určitě ne	Určitě ano
7	Spíše ne	Určitě ne	Spíše ne	Spíše ano
8	Spíše ano	Spíše ano	Spíše ano	Spíše ne
9	Spíše ne	Spíše ne	Spíše ano	Spíše ano
10	Spíše ne	Spíše ano	Spíše ano	Určitě ano
11	Spíše ano	Spíše ne	Spíše ano	Spíše ano
12	Spíše ano	Určitě ne	Spíše ano	Určitě ano
13	Spíše ano	Spíše ne	Spíše ne	Určitě ano
14	Spíše ne	Určitě ne	Spíše ne	Spíše ano
15	Spíše ano	Spíše ano	Spíše ano	Spíše ano
16	Spíše ne	Určitě ne	Určitě ne	Určitě ano
17	NA	NA	Určitě ne	Určitě ne
18	Spíše ano	Spíše ne	Spíše ne	Spíše ano
19	Spíše ano	Určitě ne	Spíše ne	Spíše ano
20	Spíše ne	Určitě ne	Spíše ne	Spíše ne
21	Spíše ano	Spíše ano	Spíše ano	Určitě ano
22	Spíše ano	Spíše ne	Spíše ano	Určitě ano
23	Spíše ne	Určitě ne	Určitě ne	Určitě ano
24	Spíše ne	Určitě ne	Spíše ne	Spíše ano
25	Spíše ne	Určitě ne	Určitě ne	Spíše ano
26	Spíše ne	Určitě ne	Určitě ne	Spíše ne
27	Spíše ne	Určitě ne	Určitě ne	Spíše ano
28	NA	NA	Určitě ne	Určitě ne
29	Spíše ano	Spíše ano	Spíše ano	Spíše ano
30	Spíše ne	Spíše ne	Spíše ne	Spíše ne
31	Spíše ne	Určitě ne	Určitě ne	Spíše ano
32	Spíše ano	Spíše ano	Určitě ano	Spíše ano
33	Spíše ne	Spíše ano	Spíše ano	Spíše ano

34	NA	NA	Určitě ne	Určitě ne
----	----	----	-----------	-----------

Tabulka 44.: Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - complete linkage

Číslo otázky	Shluk 1	Shluk 2	Shluk 3
1	Určitě ne	Spíše ne	Spíše ne
2	Spíše ne	Určitě ne	Spíše ne
3	Spíše ne	Spíše ne	Určitě ano
4	Určitě ne	Spíše ne	Spíše ne
5	Spíše ano	Spíše ne	Spíše ano
6	Spíše ne	Určitě ne	Spíše ano
7	Spíše ano	Spíše ne	Spíše ne
8	Určitě ne	Spíše ano	Určitě ano
9	Spíše ano	Spíše ano	Spíše ano
10	Spíše ne	Spíše ano	Určitě ano
11	Spíše ano	Spíše ano	Spíše ano
12	Spíše ano	Spíše ano	Spíše ne
13	Spíše ano	Spíše ne	Spíše ano
14	Spíše ne	Spíše ne	Spíše ne
15	Spíše ano	Spíše ano	Určitě ano
16	Určitě ne	Určitě ne	Určitě ne
17	NA	Určitě ne	NA
18	Určitě ano	Spíše ne	Spíše ano
19	Spíše ne	Spíše ne	Spíše ne
20	Spíše ne	Spíše ne	Spíše ne
21	Spíše ano	Spíše ano	Spíše ano
22	Spíše ne	Spíše ano	Určitě ano
23	Spíše ano	Určitě ne	Určitě ne
24	Spíše ne	Spíše ne	Spíše ano
25	Spíše ne	Určitě ne	Spíše ne
26	Spíše ne	Určitě ne	Spíše ne
27	Spíše ne	Určitě ne	Určitě ne
28	Určitě ne	Určitě ne	NA
29	Spíše ano	Spíše ano	Určitě ano
30	Spíše ne	Spíše ne	Spíše ano

31	Spíše ne	Určitě ne	Spíše ne
32	Spíše ano	Určitě ano	Určitě ano
33	Spíše ano	Spíše ano	Určitě ano
34	NA	Určitě ne	NA

Tabulka 45.: Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2
1	Spíše ne	Určitě ne
2	Spíše ne	Určitě ne
3	Spíše ano	Určitě ne
4	Určitě ne	Určitě ne
5	Spíše ano	Spíše ne
6	Spíše ne	Spíše ne
7	Spíše ne	Určitě ne
8	Spíše ne	Spíše ano
9	Spíše ano	Spíše ne
10	Spíše ano	Spíše ano
11	Spíše ne	Spíše ano
12	Spíše ne	Spíše ne
13	Spíše ano	Spíše ne
14	Spíše ne	Určitě ne
15	Spíše ano	Spíše ano
16	Určitě ne	Určitě ne
17	NA	NA
18	Spíše ano	Určitě ne
19	Spíše ne	Určitě ne
20	Spíše ne	Určitě ne
21	Spíše ano	Spíše ne
22	Spíše ne	Určitě ne
23	Určitě ne	Určitě ne
24	Spíše ne	Určitě ne
25	Spíše ano	Určitě ne
26	Určitě ne	Určitě ne

27	Určitě ne	Určitě ne
28	NA	NA
29	Určitě ano	Spíše ano
30	Určitě ne	Spíše ne
31	Spíše ne	Určitě ne
32	Určitě ano	Spíše ano
33	Určitě ano	Spíše ano
34	NA	NA

Tabulka 46.: Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - complete linkage

Číslo otázky	Shluk 1	Shluk 2
1	Spíše ne	Spíše ne
2	Určitě ne	Spíše ne
3	Spíše ne	Spíše ne
4	Spíše ne	Spíše ne
5	Spíše ne	Spíše ne
6	Určitě ne	Spíše ano
7	Spíše ne	Určitě ne
8	Spíše ano	Spíše ano
9	Spíše ano	Spíše ne
10	Spíše ano	Spíše ne
11	Spíše ano	Spíše ano
12	Spíše ano	Spíše ne
13	Spíše ne	Spíše ano
14	Spíše ne	Spíše ne
15	Spíše ano	Určitě ano
16	Určitě ne	Spíše ne
17	Určitě ne	NA
18	Spíše ne	Spíše ano
19	Spíše ne	Spíše ne
20	Spíše ne	Spíše ne
21	Spíše ano	Spíše ano
22	Spíše ano	Spíše ano

23	Určitě ne	Spíše ne
24	Spíše ne	Spíše ano
25	Určitě ne	Spíše ne
26	Určitě ne	Spíše ne
27	Určitě ne	Spíše ne
28	Určitě ne	Spíše ne
29	Spíše ano	Spíše ano
30	Spíše ne	Spíše ano
31	Určitě ne	Spíše ano
32	Určitě ano	Spíše ano
33	Spíše ano	Spíše ano
34	Určitě ne	NA

Tabulka 47.: Typičtí zástupci shluků - Apl. inf. a Inf. 2011 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Spíše ne	Určitě ne	Spíše ano
2	Spíše ne	Spíše ne	Určitě ne	Spíše ano
3	Spíše ne	Spíše ano	Určitě ne	Určitě ne
4	Určitě ne	Určitě ne	NA	Určitě ne
5	Spíše ano	Spíše ne	Spíše ano	Určitě ano
6	Spíše ano	Spíše ne	Určitě ne	Spíše ano
7	Spíše ne	Spíše ne	Určitě ne	Určitě ano
8	Určitě ano	Spíše ne	Určitě ne	NA
9	Určitě ano	Spíše ne	Určitě ne	NA
10	Spíše ne	Spíše ne	Určitě ne	Spíše ano
11	Určitě ano	Určitě ano	Spíše ne	NA
12	Spíše ne	Spíše ne	Určitě ne	NA
13	Spíše ano	Spíše ano	NA	Spíše ano
14	Určitě ne	Spíše ne	Spíše ne	Určitě ne
15	Určitě ano	Určitě ano	Spíše ano	Určitě ano
16	Spíše ne	Určitě ne	Určitě ne	Určitě ano
17	Určitě ne	NA	Určitě ne	NA
18	Určitě ano	Určitě ano	Určitě ne	Určitě ano

19	Spíše ano	Spíše ne	Určitě ne	NA
20	Spíše ne	Spíše ne	Určitě ne	NA
21	Spíše ano	Určitě ano	Určitě ano	Určitě ano
22	Spíše ano	Spíše ne	Určitě ne	NA
23	Spíše ano	Spíše ne	Určitě ne	NA
24	Spíše ano	Spíše ne	Spíše ano	NA
25	Spíše ne	Spíše ne	Určitě ne	NA
26	Spíše ne	Spíše ne	Určitě ne	NA
27	Spíše ne	Spíše ne	Určitě ne	NA
28	Určitě ne	NA	Určitě ne	NA
29	Určitě ano	Spíše ano	Určitě ano	Určitě ano
30	Spíše ne	Spíše ne	Spíše ne	NA
31	Spíše ne	Spíše ne	Určitě ne	Určitě ne
32	Určitě ano	Určitě ano	Spíše ano	Určitě ano
33	Určitě ano	Určitě ano	Spíše ano	Určitě ano
34	Určitě ne	NA	Určitě ne	NA

Tabulka 48.: Typičtí zástupci shluků - Apl. inf. a Inf. 2010 - complete linkage

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Určitě ne	Spíše ne	Určitě ne
2	Spíše ne	Určitě ne	Spíše ne	Určitě ne
3	Spíše ne	Spíše ne	Určitě ano	Určitě ne
4	Určitě ne	Určitě ne	Spíše ne	NA
5	Spíše ano	Spíše ano	Spíše ne	Spíše ano
6	Spíše ano	Určitě ne	Spíše ne	Určitě ne
7	Spíše ne	Určitě ne	Spíše ano	Určitě ne
8	Určitě ano	Určitě ano	Spíše ne	Určitě ne
9	Určitě ano	Spíše ano	Spíše ne	Určitě ne
10	Spíše ne	Určitě ano	Spíše ano	Určitě ne
11	Určitě ano	Spíše ne	Spíše ano	Spíše ne
12	Spíše ne	Spíše ano	Určitě ano	Určitě ne
13	Spíše ano	Spíše ne	Určitě ano	NA
14	Určitě ne	Spíše ne	Spíše ano	Spíše ne

15	Určitě ano	Spíše ano	Určitě ano	Spíše ano
16	Spíše ne	Určitě ne	Spíše ano	Určitě ne
17	Určitě ne	Určitě ne	Určitě ne	Určitě ne
18	Určitě ano	Spíše ano	Určitě ano	Určitě ne
19	Spíše ano	Určitě ne	Spíše ne	Určitě ne
20	Spíše ne	Určitě ne	Určitě ne	Určitě ne
21	Spíše ano	Určitě ano	Spíše ne	Určitě ano
22	Spíše ano	Spíše ne	Spíše ne	Určitě ne
23	Spíše ano	Spíše ne	Spíše ano	Určitě ne
24	Spíše ano	Spíše ne	Určitě ne	Spíše ano
25	Spíše ne	Spíše ne	Určitě ne	Určitě ne
26	Spíše ne	Spíše ne	Spíše ne	Určitě ne
27	Spíše ne	Spíše ne	Určitě ne	Určitě ne
28	Určitě ne	Určitě ne	Určitě ne	Určitě ne
29	Určitě ano	Spíše ano	Spíše ano	Určitě ano
30	Spíše ne	Spíše ne	Spíše ne	Spíše ne
31	Spíše ne	Určitě ne	Spíše ne	Určitě ne
32	Určitě ano	Spíše ano	Určitě ano	Spíše ano
33	Určitě ano	Spíše ne	Spíše ano	Spíše ano
34	Určitě ne	Určitě ne	Určitě ne	Určitě ne

Tabulka 49.: Typičtí zástupci shluků - Apl. inf. a Inf. 2010 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Spíše ne	Spíše ne	Určitě ne
2	Spíše ne	Určitě ne	Spíše ne	Určitě ne
3	Spíše ano	Spíše ne	Spíše ano	Určitě ne
4	Spíše ne	Spíše ne	Spíše ne	Určitě ne
5	Spíše ano	Spíše ne	Spíše ano	Určitě ne
6	Spíše ano	Určitě ne	Spíše ne	Určitě ne
7	Spíše ne	Spíše ne	Spíše ano	Určitě ne
8	Určitě ano	Spíše ano	Spíše ne	Určitě ne
9	Spíše ano	Spíše ano	Spíše ne	Spíše ano
10	Spíše ne	Spíše ano	Spíše ne	Spíše ano

11	Spíše ne	Spíše ano	NA	Určitě ne
12	Spíše ne	Spíše ano	Spíše ano	Spíše ano
13	Spíše ano	Spíše ne	Spíše ano	Určitě ne
14	Spíše ne	Spíše ne	Spíše ne	Určitě ne
15	Spíše ano	Spíše ano	Spíše ano	Určitě ne
16	Spíše ne	Určitě ne	Spíše ne	Určitě ne
17	NA	Určitě ne	Spíše ne	Určitě ne
18	Spíše ano	Spíše ne	Spíše ne	Spíše ano
19	Spíše ne	Spíše ne	Spíše ne	Určitě ne
20	Spíše ne	Spíše ne	Spíše ne	Určitě ne
21	Spíše ano	Spíše ano	Spíše ano	Určitě ne
22	Spíše ano	Spíše ano	Spíše ano	Určitě ne
23	Spíše ne	Určitě ne	Spíše ano	Určitě ne
24	Spíše ano	Spíše ne	Spíše ano	Určitě ne
25	Spíše ne	Určitě ne	Spíše ne	Určitě ne
26	Spíše ne	Určitě ne	Spíše ne	Určitě ne
27	Určitě ne	Určitě ne	Spíše ne	Určitě ne
28	NA	Určitě ne	Spíše ne	Určitě ne
29	Spíše ano	Spíše ano	Spíše ano	Určitě ne
30	Spíše ne	Spíše ne	Spíše ano	Určitě ne
31	Určitě ne	Určitě ne	Spíše ne	Určitě ne
32	Určitě ano	Určitě ano	Spíše ano	Spíše ne
33	Spíše ano	Spíše ano	Spíše ano	Určitě ne
34	NA	Určitě ne	Spíše ne	Určitě ne

Tabulka 50.: Typičtí zástupci shluků - Apl. inf. a Inf. 2010 a 2011 - complete linkage

Číslo otázky	Shluk 1	Shluk 2	Shluk 3
1	Spíše ne	Spíše ne	Určitě ne
2	Spíše ne	Spíše ne	Určitě ne
3	Spíše ano	Spíše ne	Určitě ne
4	Spíše ne	Spíše ne	Určitě ne
5	Spíše ano	Spíše ano	Spíše ne
6	Spíše ne	Spíše ano	Spíše ne

7	Určitě ne	Spíše ano	Určitě ne
8	Spíše ano	Spíše ano	Spíše ano
9	Spíše ne	Spíše ano	Spíše ne
10	Spíše ano	Spíše ano	Spíše ano
11	Spíše ne	Určitě ano	Spíše ano
12	Spíše ano	Spíše ne	Spíše ne
13	Spíše ano	Určitě ano	Spíše ne
14	Spíše ne	Spíše ne	Určitě ne
15	Spíše ano	Spíše ano	Spíše ano
16	Určitě ne	Určitě ne	Určitě ne
17	Spíše ne	NA	NA
18	Spíše ano	Spíše ano	Určitě ne
19	Spíše ne	Spíše ne	Určitě ne
20	Spíše ne	Spíše ne	Určitě ne
21	Spíše ano	Spíše ano	Spíše ne
22	Spíše ne	Spíše ano	Určitě ne
23	Spíše ne	Spíše ano	Určitě ne
24	Spíše ano	Spíše ano	Určitě ne
25	Spíše ne	Spíše ne	Určitě ne
26	Spíše ne	Spíše ne	Určitě ne
27	Spíše ne	Spíše ne	Určitě ne
28	Spíše ne	Spíše ne	NA
29	Spíše ano	Určitě ano	Spíše ano
30	Spíše ne	Spíše ano	Spíše ne
31	Spíše ne	NA	Určitě ne
32	Spíše ano	Určitě ano	Spíše ano
33	Spíše ano	Určitě ano	Spíše ano
34	Spíše ne	NA	NA

Tabulka 51.: Typičtí zástupci shluků - Apl. inf. a Inf. 2010 a 2011 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3
1	Určitě ne	Určitě ne	Spíše ne
2	Určitě ne	Určitě ne	Určitě ne

3	Určitě ano	Určitě ne	Spíše ne
4	Určitě ne	Určitě ne	Určitě ne
5	Určitě ano	Určitě ne	Spíše ne
6	Určitě ano	Určitě ne	Určitě ne
7	Spíše ne	Spíše ne	Určitě ne
8	Určitě ne	Určitě ne	Určitě ano
9	Spíše ano	Určitě ne	Spíše ne
10	Určitě ne	Určitě ne	Určitě ne
11	Určitě ano	Spíše ano	Spíše ne
12	Spíše ne	Spíše ne	Spíše ano
13	Určitě ano	Spíše ne	Spíše ano
14	Spíše ne	Spíše ne	Určitě ne
15	Určitě ano	Určitě ano	Spíše ano
16	Určitě ne	Určitě ne	Určitě ne
17	NA	Spíše ne	NA
18	Určitě ano	Určitě ne	Určitě ne
19	Určitě ano	Spíše ne	Spíše ano
20	Určitě ne	Spíše ne	Spíše ano
21	Určitě ano	Určitě ano	Spíše ano
22	Určitě ano	Určitě ano	Spíše ano
23	Určitě ano	Spíše ne	Spíše ano
24	Určitě ano	Spíše ne	Spíše ne
25	Určitě ne	Určitě ano	Spíše ne
26	Určitě ne	Určitě ne	Spíše ne
27	Určitě ne	Určitě ne	Spíše ne
28	NA	Určitě ne	Určitě ne
29	Určitě ano	Spíše ano	Spíše ne
30	Určitě ano	Určitě ano	Určitě ne
31	Určitě ne	Spíše ne	Určitě ne
32	Určitě ano	Určitě ano	Určitě ano
33	Určitě ano	Spíše ano	Určitě ano
34	NA	Určitě ne	Určitě ne

Tabulka 52.: Typičtí zástupci shluků - 2011 - complete linkage

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Určitě ne	Spíše ne	Určitě ne	Určitě ne
2	Určitě ne	Spíše ne	Určitě ne	Určitě ne
3	Určitě ne	Spíše ne	Určitě ano	Určitě ne
4	Určitě ne	Spíše ne	Určitě ne	Určitě ne
5	Spíše ano	Spíše ne	Určitě ano	Určitě ne
6	Určitě ne	Spíše ne	Určitě ano	Určitě ne
7	Určitě ne	Spíše ne	Určitě ne	Spíše ne
8	Určitě ano	Určitě ano	Určitě ne	Určitě ne
9	Určitě ne	Určitě ano	Určitě ne	Určitě ne
10	Spíše ne	Určitě ano	Určitě ne	Určitě ne
11	Spíše ne	Spíše ano	Spíše ne	Spíše ano
12	Spíše ne	Určitě ano	Určitě ne	Spíše ne
13	Spíše ne	Spíše ano	Určitě ano	Spíše ne
14	Spíše ne	Spíše ne	Určitě ne	Spíše ne
15	Určitě ano	Spíše ano	Spíše ano	Určitě ano
16	Určitě ne	Určitě ne	Určitě ne	Určitě ne
17	NA	NA	Určitě ne	Spíše ne
18	Spíše ano	Určitě ano	Spíše ne	Určitě ne
19	Spíše ne	Určitě ano	Spíše ne	Spíše ne
20	Určitě ne	Spíše ano	Spíše ano	Spíše ne
21	Určitě ano	Spíše ano	Spíše ano	Určitě ano
22	Spíše ne	Spíše ano	Určitě ano	Určitě ano
23	Určitě ne	Určitě ano	Určitě ano	Spíše ne
24	Určitě ano	Spíše ano	Spíše ano	Spíše ne
25	Určitě ne	Spíše ne	Určitě ne	Určitě ano
26	Určitě ne	Spíše ne	Určitě ne	Určitě ne
27	Určitě ne	Určitě ne	Určitě ne	Určitě ne
28	Určitě ne	NA	Určitě ne	Určitě ne
29	Spíše ano	Určitě ano	Spíše ano	Spíše ano
30	Spíše ne	Spíše ne	Určitě ne	Určitě ano
31	Spíše ne	Spíše ne	Určitě ne	Spíše ne
32	Určitě ano	Určitě ano	Spíše ano	Určitě ano
33	Spíše ano	Spíše ano	Určitě ano	Spíše ano

34	NA	NA	Určitě ne	Určitě ne
----	----	----	-----------	-----------

Tabulka 53.: Typičtí zástupci shluků - 2011 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3
1	Spíše ne	Spíše ne	Spíše ano
2	Spíše ne	Určitě ne	Spíše ano
3	Spíše ano	Určitě ne	Určitě ano
4	Spíše ne	Spíše ne	NA
5	Spíše ano	Spíše ne	Určitě ano
6	Spíše ne	Spíše ne	Spíše ano
7	Určitě ne	Spíše ne	Spíše ano
8	Spíše ne	Spíše ano	NA
9	Spíše ne	Spíše ne	Spíše ano
10	Spíše ne	Spíše ano	NA
11	Spíše ano	Spíše ne	NA
12	Spíše ne	Spíše ano	NA
13	Spíše ano	NA	NA
14	Spíše ne	Spíše ne	NA
15	Spíše ano	Spíše ne	NA
16	Spíše ne	Určitě ne	NA
17	NA	NA	NA
18	Spíše ne	Spíše ne	Určitě ano
19	Spíše ne	Spíše ne	Určitě ano
20	Spíše ne	Spíše ne	Spíše ne
21	Určitě ano	Spíše ano	Určitě ano
22	Spíše ano	Spíše ne	Určitě ano
23	Spíše ano	Spíše ne	Spíše ano
24	Spíše ano	Spíše ne	Určitě ano
25	Spíše ne	Určitě ne	Spíše ne
26	Spíše ne	Určitě ne	Spíše ne
27	Spíše ne	Určitě ne	Spíše ne
28	Spíše ne	NA	NA
29	Určitě ano	Spíše ne	NA

30	Určitě ano	Určitě ne	NA
31	Určitě ne	Spíše ne	NA
32	Spíše ano	Spíše ne	Určitě ano
33	Určitě ano	Spíše ne	NA
34	NA	NA	NA

Tabulka 54.: Typičtí zástupci shluků s obory - 2011 - complete linkage

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Určitě ne	Spíše ne	Spíše ne	Spíše ne
2	Určitě ne	Spíše ne	Spíše ne	Spíše ne
3	Spíše ne	Spíše ano	Spíše ne	Spíše ano
4	Určitě ne	Spíše ne	Spíše ne	Spíše ne
5	NA	Spíše ano	Spíše ano	Spíše ano
6	Spíše ne	Spíše ne	Spíše ne	Spíše ne
7	Spíše ano	Určitě ne	Spíše ne	Určitě ne
8	Určitě ne	Spíše ne	Spíše ne	Spíše ano
9	Spíše ne	Spíše ne	Spíše ne	Spíše ne
10	Spíše ne	Spíše ne	Spíše ne	Spíše ano
11	Spíše ano	Spíše ano	Spíše ne	Spíše ne
12	Spíše ano	Spíše ne	Spíše ne	Spíše ano
13	Spíše ano	Spíše ano	Spíše ne	Spíše ano
14	Spíše ne	Spíše ne	Spíše ne	Spíše ne
15	Spíše ano	Spíše ano	Spíše ano	Spíše ano
16	Spíše ne	Spíše ne	Určitě ne	Určitě ne
17	NA	NA	NA	Spíše ne
18	Určitě ano	Spíše ne	Spíše ano	Spíše ano
19	Spíše ne	Spíše ne	Spíše ne	Spíše ne
20	Spíše ne	Spíše ne	Spíše ne	Spíše ne
21	Spíše ano	Určitě ano	Spíše ano	Spíše ano
22	Spíše ano	Spíše ano	Spíše ne	Spíše ne
23	Spíše ne	Spíše ano	Spíše ne	Spíše ne
24	Spíše ano	Spíše ano	Spíše ne	Spíše ano
25	Určitě ne	Spíše ne	Určitě ne	Spíše ne

26	Určitě ne	Spíše ne	Spíše ne	Spíše ne
27	Určitě ne	Spíše ne	Určitě ne	Spíše ne
28	NA	Spíše ne	NA	Spíše ne
29	Spíše ano	Určitě ano	Spíše ne	Spíše ano
30	Spíše ne	Určitě ano	Spíše ne	Spíše ne
31	Spíše ne	Určitě ne	Spíše ne	Spíše ne
32	Určitě ano	Spíše ano	Spíše ano	Spíše ano
33	Určitě ano	Určitě ano	Spíše ano	Spíše ano
34	NA	NA	NA	Spíše ne

Tabulka 55.: Typičtí zástupci shluků s obory - 2011 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Spíše ne	Určitě ne	Určitě ne
2	Spíše ne	Spíše ne	Určitě ne	Spíše ne
3	Spíše ne	Spíše ano	Spíše ne	Spíše ne
4	Spíše ne	Určitě ne	Určitě ne	Určitě ne
5	Spíše ne	Spíše ne	Spíše ano	Určitě ne
6	Spíše ano	Spíše ne	Spíše ano	Určitě ne
7	Spíše ne	Spíše ne	Spíše ne	Spíše ne
8	Spíše ne	Spíše ne	Určitě ne	Spíše ne
9	Spíše ne	Spíše ne	Určitě ne	Spíše ne
10	Spíše ne	Spíše ne	Určitě ne	Spíše ne
11	Spíše ano	Určitě ano	Spíše ne	Spíše ne
12	Spíše ano	Spíše ne	Spíše ano	Spíše ne
13	Určitě ano	Spíše ano	Spíše ano	Spíše ano
14	Spíše ne	Spíše ne	Spíše ne	Spíše ne
15	Určitě ano	Určitě ano	Spíše ano	Spíše ano
16	Určitě ne	Určitě ne	Určitě ne	Určitě ano
17	NA	NA	NA	NA
18	Spíše ne	Určitě ano	Spíše ne	Určitě ano
19	Spíše ano	Spíše ne	Spíše ne	Určitě ano
20	Spíše ne	Spíše ne	Určitě ne	Určitě ne
21	Spíše ano	Určitě ano	Určitě ano	Spíše ano

22	Určitě ano	Spíše ne	Spíše ano	Spíše ano
23	Spíše ano	Spíše ne	Spíše ano	Spíše ano
24	Spíše ano	Spíše ne	Spíše ano	Spíše ne
25	Určitě ne	Spíše ne	Určitě ne	Určitě ne
26	Určitě ne	Spíše ne	Určitě ne	Určitě ne
27	Určitě ne	Spíše ne	Určitě ne	Spíše ano
28	NA	NA	Určitě ne	Určitě ne
29	Spíše ano	Spíše ano	Spíše ano	Spíše ne
30	Spíše ano	Spíše ne	Určitě ne	Spíše ne
31	Určitě ne	Spíše ne	Spíše ne	Spíše ne
32	Určitě ano	Určitě ano	Spíše ano	Spíše ano
33	Spíše ano	Určitě ano	Určitě ano	Spíše ano
34	NA	NA	NA	NA

Tabulka 56.: Typičtí zástupci shluků - 2010 - complete linkage

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Spíše ne	Spíše ne	Určitě ne
2	Spíše ne	Určitě ne	Určitě ne	Spíše ano
3	Spíše ne	Určitě ne	Spíše ne	Určitě ne
4	Spíše ne	Určitě ne	Určitě ne	Určitě ne
5	Spíše ne	Spíše ano	Určitě ano	Spíše ano
6	Spíše ano	Spíše ano	Určitě ano	Spíše ne
7	Spíše ne	Spíše ano	Spíše ano	Spíše ne
8	Spíše ne	Určitě ne	Spíše ne	Spíše ne
9	Spíše ne	Spíše ano	Určitě ne	Spíše ano
10	Spíše ne	Spíše ano	Určitě ne	Určitě ne
11	Spíše ano	Spíše ano	Určitě ano	Určitě ano
12	Spíše ano	Spíše ano	Určitě ano	Spíše ano
13	Určitě ano	Spíše ano	Určitě ano	Určitě ano
14	Spíše ne	Určitě ne	Určitě ne	Určitě ne
15	Určitě ano	Určitě ano	Určitě ano	Určitě ano
16	Určitě ne	Určitě ne	Určitě ne	Určitě ne
17	NA	Určitě ne	NA	NA

18	Spíše ne	Spíše ano	Určitě ano	Spíše ano
19	Spíše ano	Určitě ne	Určitě ano	Určitě ano
20	Spíše ne	Určitě ne	Spíše ne	Určitě ne
21	Spíše ano	Určitě ano	Určitě ano	Spíše ne
22	Určitě ano	Spíše ano	Určitě ano	Určitě ano
23	Spíše ano	Určitě ne	Určitě ano	Určitě ano
24	Spíše ano	Určitě ne	Určitě ano	Určitě ano
25	Určitě ne	Spíše ne	Spíše ne	Určitě ne
26	Určitě ne	Určitě ne	Určitě ne	Určitě ne
27	Určitě ne	Určitě ne	Určitě ne	Určitě ne
28	NA	Určitě ne	Určitě ne	NA
29	Spíše ano	Určitě ano	Určitě ano	Spíše ne
30	Spíše ano	Spíše ano	Spíše ano	Určitě ano
31	Určitě ne	Spíše ne	Určitě ne	Určitě ne
32	Určitě ano	Určitě ano	Určitě ano	Určitě ano
33	Spíše ano	Určitě ano	Určitě ano	Spíše ano
34	NA	Určitě ne	NA	NA

Tabulka 57.: Typičtí zástupci shluků - 2010 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	NA	Určitě ne	Určitě ne	Určitě ne
2	Spíše ano	Spíše ne	Určitě ne	Spíše ne
3	Spíše ne	Určitě ano	Určitě ne	Určitě ne
4	Určitě ne	Spíše ne	Určitě ne	Spíše ano
5	Spíše ano	Spíše ano	Spíše ne	Spíše ano
6	Spíše ano	Spíše ano	Spíše ne	Určitě ano
7	Spíše ne	Spíše ano	Spíše ne	Určitě ne
8	Spíše ano	Určitě ne	Spíše ano	Spíše ano
9	Spíše ano	Určitě ne	Spíše ano	Určitě ano
10	NA	Spíše ne	Spíše ano	Určitě ne
11	Spíše ano	Spíše ano	Spíše ano	Určitě ano
12	Spíše ne	Spíše ano	Spíše ano	Spíše ne
13	Spíše ano	Spíše ano	Spíše ano	Spíše ano

14	NA	Určitě ne	Spíše ne	Určitě ne
15	Určitě ano	Určitě ano	Určitě ano	Určitě ano
16	Určitě ne	Určitě ne	Určitě ne	Určitě ne
17	NA	NA	Určitě ano	NA
18	Spíše ano	Spíše ne	Určitě ano	Spíše ano
19	Spíše ano	Spíše ano	NA	Spíše ano
20	Spíše ne	Spíše ne	NA	Určitě ne
21	Spíše ano	Spíše ano	NA	Spíše ano
22	Spíše ano	Určitě ano	NA	Spíše ano
23	Spíše ano	Spíše ne	NA	Určitě ne
24	Spíše ano	Spíše ano	NA	Spíše ano
25	Spíše ne	Určitě ne	NA	Určitě ne
26	Určitě ne	Určitě ne	NA	Spíše ano
27	Určitě ne	Určitě ne	NA	Určitě ano
28	NA	NA	NA	NA
29	Určitě ano	Spíše ano	NA	Určitě ano
30	Spíše ano	Spíše ano	NA	Určitě ano
31	Určitě ne	Určitě ne	NA	Určitě ne
32	Určitě ano	Určitě ano	NA	Určitě ano
33	Spíše ano	Spíše ano	NA	Určitě ano
34	NA	NA	NA	NA

Tabulka 58.: Typičtí zástupci shluků s obory - 2010 - complete linkage

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
1	Určitě ne	Spíše ne	Spíše ne	Určitě ne	NA
2	NA	Spíše ne	Spíše ne	Určitě ne	Spíše ano
3	NA	Spíše ne	Spíše ano	Určitě ne	Spíše ne
4	NA	Určitě ne	Spíše ne	Určitě ne	Určitě ne
5	NA	Spíše ne	Spíše ano	Spíše ne	Spíše ano
6	NA	Určitě ne	Spíše ne	Spíše ne	Spíše ano
7	Určitě ano	Spíše ne	Spíše ano	Spíše ne	Spíše ne
8	Určitě ne	Určitě ne	Spíše ne	Spíše ano	Spíše ano
9	Určitě ne	Spíše ne	Spíše ne	Spíše ano	Spíše ano

10	Určitě ne	Určitě ne	Spíše ne	Spíše ano	NA
11	Spíše ano	Spíše ano	NA	Spíše ano	Spíše ano
12	Spíše ne	Spíše ne	Spíše ano	Spíše ano	Spíše ne
13	Spíše ano	Spíše ne	Spíše ano	Spíše ano	Spíše ano
14	Určitě ne	Určitě ne	Spíše ne	Spíše ne	NA
15	Určitě ano	Spíše ne	Spíše ano	Určitě ano	Určitě ano
16	Určitě ne	Určitě ne	Spíše ne	Určitě ne	Určitě ne
17	NA	Určitě ne	Spíše ne	Určitě ano	NA
18	Spíše ano	Spíše ne	Spíše ne	Určitě ano	Spíše ano
19	Spíše ano	Spíše ne	Spíše ne	NA	Spíše ano
20	Spíše ne	Spíše ne	Spíše ne	NA	Spíše ne
21	Spíše ano	Spíše ano	Spíše ano	NA	Spíše ano
22	Spíše ano	Spíše ne	Spíše ano	NA	Spíše ano
23	Určitě ne	Spíše ne	Spíše ano	NA	Spíše ano
24	Spíše ano	Určitě ne	Spíše ano	NA	Spíše ano
25	Určitě ne	Určitě ne	Spíše ne	NA	Spíše ne
26	Určitě ne	Určitě ne	Spíše ne	NA	Určitě ne
27	Určitě ne	Určitě ne	Spíše ne	NA	Určitě ne
28	NA	Spíše ne	Spíše ne	NA	NA
29	Určitě ano	Spíše ano	Spíše ano	NA	Určitě ano
30	Spíše ano	Spíše ano	Spíše ano	NA	Spíše ano
31	Určitě ne	Určitě ne	Spíše ne	NA	Určitě ne
32	Určitě ano	Spíše ano	Spíše ano	NA	Určitě ano
33	Určitě ano	Spíše ano	Spíše ano	NA	Spíše ano
34	NA	Určitě ne	Spíše ne	NA	NA

Tabulka 59.: Typičtí zástupci shluků s obory - 2010 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Spíše ne	Spíše ne	Spíše ne
2	Určitě ano	Spíše ne	Spíše ne	Spíše ano
3	Spíše ano	Určitě ne	Spíše ne	Spíše ano
4	Spíše ne	Určitě ne	Spíše ne	Určitě ne
5	Spíše ne	Určitě ne	Spíše ano	Určitě ne

6	Spíše ano	Určitě ne	Spíše ano	Spíše ne
7	Určitě ano	Spíše ne	Spíše ano	Určitě ne
8	Určitě ne	Určitě ne	Spíše ano	Určitě ne
9	Spíše ano	Určitě ne	Spíše ano	Spíše ano
10	Spíše ne	Určitě ne	Spíše ano	Spíše ano
11	Spíše ano	Spíše ano	Určitě ano	Spíše ano
12	Spíše ano	Spíše ano	Spíše ne	Spíše ano
13	Spíše ne	Spíše ne	Určitě ano	Spíše ano
14	Spíše ne	Určitě ne	Spíše ne	Spíše ne
15	Spíše ano	Spíše ano	Spíše ano	Určitě ano
16	Spíše ne	Určitě ne	Určitě ne	Spíše ne
17	NA	NA	NA	NA
18	Spíše ano	Určitě ne	Spíše ano	Spíše ne
19	Určitě ano	Určitě ne	Spíše ne	Spíše ne
20	Spíše ne	Určitě ne	Spíše ne	Spíše ne
21	Spíše ne	Určitě ano	Spíše ano	Spíše ano
22	Určitě ano	Určitě ne	Spíše ano	Spíše ano
23	Spíše ne	Určitě ne	Spíše ano	Spíše ano
24	Spíše ne	Spíše ano	Spíše ano	Spíše ne
25	Spíše ne	Určitě ne	Spíše ne	Spíše ne
26	Spíše ne	Určitě ne	Spíše ne	Spíše ne
27	Spíše ne	Určitě ne	Spíše ne	Spíše ne
28	Určitě ne	NA	Spíše ne	Spíše ne
29	Spíše ano	Spíše ne	Určitě ano	Spíše ano
30	Spíše ne	Spíše ne	Spíše ano	Spíše ne
31	Spíše ne	Určitě ne	NA	Spíše ne
32	Určitě ano	Určitě ano	Určitě ano	Určitě ano
33	Spíše ano	Určitě ano	Určitě ano	Určitě ano
34	NA	NA	NA	NA

Tabulka 60.: Typičtí zástupci shluků - 2010 a 2011 - complete linkage

Č. ot.	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5	Shluk 6
1	Spíše ne	Spíše ne	Spíše ne	Spíše ne	Spíše ano	Určitě ne

2	Spíše ne	Určitě ne	Určitě ne	Spíše ne	Určitě ne	Určitě ne
3	Spíše ne	Spíše ne	Spíše ne	Určitě ne	Určitě ne	Určitě ne
4	Spíše ne	Spíše ne	Určitě ne	Určitě ne	Určitě ne	Určitě ne
5	Spíše ano	Spíše ano	Určitě ano	NA	Určitě ne	Určitě ne
6	Spíše ano	Spíše ano	Určitě ano	Určitě ne	Spíše ano	Určitě ne
7	Spíše ano	Určitě ne	Spíše ano	Určitě ne	Spíše ne	Spíše ne
8	Spíše ano	Určitě ano	Spíše ne	Spíše ano	Určitě ano	Spíše ano
9	Spíše ano	Určitě ne	Určitě ne	Spíše ne	Spíše ano	Spíše ne
10	Spíše ano	Spíše ne	Určitě ne	Určitě ne	Určitě ne	Určitě ano
11	Určitě ano	Spíše ano	Určitě ano	Určitě ne	Spíše ano	Spíše ne
12	Spíše ne	Určitě ne	Určitě ano	Spíše ne	Určitě ne	Spíše ano
13	Určitě ano	Spíše ano	Určitě ano	Spíše ne	Spíše ano	Spíše ne
14	Spíše ne	Spíše ne	Určitě ne	Spíše ne	Určitě ne	Určitě ne
15	Spíše ano	Určitě ano	Určitě ano	Určitě ne	Spíše ano	Určitě ano
16	Určitě ne	Spíše ano	Určitě ne	Spíše ano	Určitě ne	Určitě ne
17	NA	NA	NA	Určitě ne	NA	Určitě ne
18	Spíše ano	Určitě ano	Určitě ano	Spíše ano	Určitě ano	Určitě ne
19	Spíše ne	Spíše ne	Určitě ano	Určitě ano	Spíše ano	Určitě ne
20	Spíše ne	Určitě ne	Spíše ne	Určitě ne	Spíše ano	Určitě ne
21	Spíše ano	Určitě ano	Určitě ano	Spíše ne	Spíše ano	Určitě ano
22	Spíše ano	Určitě ano	Určitě ano	Určitě ne	Určitě ne	Určitě ne
23	Spíše ano	Určitě ano	Určitě ano	Určitě ne	Určitě ne	Určitě ne
24	Spíše ano	Spíše ano	Určitě ano	Spíše ne	Určitě ne	Určitě ne
25	Spíše ne	Spíše ne	Spíše ne	Určitě ne	Určitě ne	Spíše ne
26	Spíše ne	Spíše ne	Určitě ne	Spíše ano	Určitě ne	Určitě ne
27	Spíše ne	Spíše ne	Určitě ne	Určitě ne	Určitě ne	Určitě ne
28	Spíše ne	NA	Určitě ne	Určitě ne	NA	Určitě ne
29	Určitě ano	Spíše ano	Určitě ano	Spíše ne	Spíše ne	Určitě ano
30	Spíše ano	Určitě ano	Spíše ano	Určitě ne	Určitě ne	Spíše ne
31	NA	Určitě ne	Určitě ne	Určitě ne	Spíše ne	Určitě ne
32	Určitě ano	Určitě ano	Určitě ano	Určitě ano	Určitě ano	Určitě ano
33	Určitě ano	Spíše ano	Určitě ano	Určitě ano	Spíše ano	Určitě ano
34	NA	NA	NA	Určitě ne	NA	Určitě ne

Tabulka 61.: Typičtí zástupci shluků - 2010 a 2011 - Wardova metoda

Číslo otázky	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	Spíše ne	Určitě ne	NA	Spíše ne
2	Spíše ne	Určitě ne	Spíše ano	Spíše ano
3	Spíše ano	Spíše ne	Spíše ne	Spíše ne
4	Spíše ne	Určitě ne	Určitě ne	Spíše ne
5	Spíše ano	Spíše ano	Spíše ano	Spíše ano
6	Spíše ne	Určitě ne	Spíše ano	Spíše ano
7	Určitě ne	Určitě ne	Spíše ne	Spíše ano
8	Spíše ne	Spíše ano	Spíše ano	Určitě ano
9	Spíše ne	Spíše ano	Spíše ano	Spíše ano
10	Spíše ne	Spíše ano	NA	Spíše ano
11	Spíše ano	Spíše ne	Spíše ano	Spíše ne
12	Spíše ne	Spíše ne	Spíše ne	Spíše ano
13	Spíše ano	Spíše ano	Spíše ano	Spíše ano
14	Spíše ne	Spíše ne	NA	Spíše ne
15	Spíše ano	Spíše ano	Určitě ano	Spíše ano
16	Spíše ne	Spíše ne	Určitě ne	NA
17	NA	NA	NA	NA
18	Spíše ne	Určitě ne	Spíše ano	Spíše ano
19	Spíše ne	Určitě ne	Spíše ano	Spíše ne
20	Spíše ne	Určitě ne	Spíše ne	Spíše ne
21	Určitě ano	Určitě ano	Spíše ano	Určitě ano
22	Spíše ano	Spíše ne	Spíše ano	Spíše ano
23	Spíše ano	Spíše ne	Spíše ano	Určitě ne
24	Spíše ano	Určitě ano	Spíše ano	Určitě ne
25	Spíše ne	Určitě ne	Spíše ne	Spíše ano
26	Spíše ne	Určitě ne	Určitě ne	Spíše ne
27	Spíše ne	Určitě ne	Určitě ne	Spíše ne
28	Spíše ne	NA	NA	NA
29	Určitě ano	Určitě ano	Určitě ano	Určitě ano
30	Určitě ano	Spíše ne	Spíše ano	Spíše ne
31	Určitě ne	Určitě ne	Určitě ne	Spíše ano
32	Spíše ano	Spíše ano	Určitě ano	Spíše ano
33	Určitě ano	Spíše ano	Spíše ano	Určitě ano

34	NA	NA	NA	NA
----	----	----	----	----

Tabulka 62.: Typičtí zástupci shluků s obory - 2010 a 2011 - complete linkage

Č. ot.	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5	Shluk 6
1	Spíše ne	Spíše ne	Určitě ne	Určitě ne	Určitě ne	Spíše ne
2	Spíše ne	Určitě ne	Určitě ne	Určitě ne	Spíše ne	Určitě ne
3	Spíše ano	Spíše ne	Určitě ne	Určitě ne	Spíše ne	Určitě ne
4	Spíše ne	Určitě ne	Určitě ne	Určitě ne	Spíše ne	Spíše ne
5	Spíše ano	Spíše ano	Spíše ne	Spíše ne	Spíše ne	Spíše ne
6	Spíše ne	Spíše ano	Spíše ne	Spíše ne	Spíše ne	Určitě ne
7	Určitě ne	Spíše ne	Spíše ne	Určitě ne	Spíše ano	NA
8	Spíše ne	Určitě ne	Spíše ano	Spíše ano	Určitě ne	Určitě ano
9	Spíše ne	Určitě ne	Spíše ano	Určitě ne	Spíše ne	Spíše ano
10	Spíše ne	Spíše ano	Spíše ano	Určitě ne	Spíše ne	Určitě ano
11	Spíše ano	Spíše ne	Spíše ano	Spíše ne	Spíše ne	Spíše ne
12	Spíše ne	Spíše ne	Spíše ano	Spíše ne	Spíše ano	NA
13	Spíše ano	Spíše ano	Spíše ano	Spíše ano	Spíše ano	NA
14	Spíše ne	Spíše ne	Spíše ne	Určitě ne	Spíše ne	NA
15	Spíše ano	Spíše ano	Určitě ano	Spíše ano	Spíše ano	Spíše ano
16	Spíše ne	Spíše ano	Určitě ne	Určitě ne	Určitě ne	Určitě ne
17	NA	NA	Určitě ano	Určitě ne	NA	NA
18	Spíše ne	Spíše ano	Určitě ano	Určitě ne	Spíše ano	NA
19	Spíše ne	Určitě ne	NA	Určitě ne	Spíše ne	NA
20	Spíše ne	Spíše ne	NA	Určitě ne	Spíše ne	NA
21	Určitě ano	Spíše ano	NA	Spíše ano	Spíše ano	NA
22	Spíše ano	Spíše ano	NA	Určitě ne	Spíše ne	NA
23	Spíše ano	Spíše ne	NA	Určitě ne	Spíše ano	NA
24	Spíše ano	Spíše ne	NA	Určitě ne	Spíše ne	NA
25	Spíše ne	Spíše ne	NA	Určitě ne	Spíše ne	NA
26	Spíše ne	Spíše ne	NA	Určitě ne	Spíše ne	Určitě ano
27	Spíše ne	Určitě ne	NA	Určitě ne	Spíše ne	NA
28	Spíše ne	NA	NA	Určitě ne	Spíše ne	NA
29	Určitě ano	Spíše ano	NA	Spíše ne	Spíše ano	Spíše ano

30	Určitě ano	Spíše ne	NA	Spíše ano	Spíše ne	Spíše ne
31	Určitě ne	Spíše ne	NA	Určitě ne	Spíše ne	Určitě ne
32	Spíše ano	Spíše ano	NA	Spíše ano	Určitě ano	Spíše ano
33	Určitě ano	Spíše ano	NA	Spíše ano	Spíše ano	Spíše ano
34	NA	NA	NA	Určitě ne	Určitě ne	NA

Tabulka 63.: Typičtí zástupci shluků s obory - 2010 a 2011 - Wardova metoda