Posudek vedoucího bakalářské práce

Název: Archetypální analýza jako segmentační nástroj
Autor: Karolína Rezková

Shrnutí obsahu práce

Předložená práce, pokud jde o její strukturu, je rozdělena na úvod, tři kapitoly a v závěru obsahuje shrnutí a seznam tabulek, obrázků a použité literatury. K dispozici je rovněž nosič CD, na kterém jsou obrázky ve formátu PS použité v práci a data pro praktickou část. Kapitola 1 přináší podrobnější důkaz základní věty o poloze archetypů, která byla v pilotním článku [Cutler a Breiman, 1994] uvedena jen s velmi krátkým důkazem. Kapitola 2 popisuje některé metody pro prvotní zpracování dat a představuje dataset o nákupním chování spotřebitelů, analyzovaný v části 3. Část 3 nejprve rozebírá 2 přístupy, jak lze na základě archetypů segmentovat data a ukazuje, že tyto přístupy nemusí vést ke stejným výsledkům. Dále autorka aplikuje segmentaci na reálná data o analýze spotřebitelů. Závěr shrnuje přínosy autorky k uvedené problematice.

Celkové hodnocení práce

Téma práce. Tématem práce byla archetypální analýza a její využití pro segmentace. Téma zahrnovalo znalosti z konvexní optimalizace v teoretické části a statistické zpracování dat v části praktické. Souvislost mezi archetypy a jejich použitím pro segmentační účely nebyla doposud podrobněji zkoumána, literatura v této oblasti se zaměřuje např. na robustifikaci algoritmu vzhledem k odlehlým pozorováním a na efektivnější algoritmizaci problému pro dosažení větší časové úspory výpočtu. Autorka porovnala navržené způsoby využití archetypů pro segmentace, samostatně poukázala na jejich rozdíly a svoje zjištění zčásti rovněž prověřila na reálných datech. Proto lze konstatovat, že splnila požadavky dané zadáním.

Vlastní příspěvek. Práce obsahuje vlastní příspěvek autorky ve formě rozšířeného a podrobného důkazu o poloze archetypů, který byl v pilotním článku [Cutler a Breiman, 1994] pouze velmi stručně naznačen. Vlastním přínosem jsou rovněž ukázkové příklady v částech 3.1.1. a 3.2.2 a zpracování reálných dat v kapitolách 2 a 3. V předložené práci je dostatečné množství vlastního materiálu.

Matematická úroveň. Matematická úroveň práce je podprůměrná, kvůli nepřesné definici 1 a četnějším výtkám ke značení a formulacím v kapitole 2, viz připomínky.

Práce se zdroji. V seznamu literatury u článku [Eugster a Leisch, 2009] chybí stránkový rozsah.

Formální úprava. Práce vzhledem k rozsahu obsahuje přiměřené množství formálních nedostatků.

Připomínky a otázky

Formální nedostatky:

- str.2[1]: "...této práce je představit..."
- str.3, z definice 1 bych oddělil text, začínající "Poznamenejme..."
- Vektory jsou v kapitolách 2 a 3 značeny tučně, kdežto v kapitole 1 netučně.
- V názvu kapitoly 2 bych místo termínu "surových"použil termín "zdrojových".
- V úvodu kapitoly 2 se hned třikrát v rychlém sledu po sobě vyskytují obdoby slova "zpracovat".
- str.16$_1$: "Za nevýhodu lze..."
- str.17$_{13}$: "...skutečně mohou být..."
- str.19$_4$: "...vzhledem k maximální hodnotě..."

Matematické nedostatky:

- str.3: Definice 1 není zcela v pořádku.

- str.10[11]: Hodnoty náhodné veličiny jsou reálná čísla, která netvoří náhodný výběr. Autorka měla zřejmě na mysli realizaci náhodného výběru.
- str.10: V definici výběrové korelační matice R chybí vysvětlení u s_{ij}.
- str.10: U vzorce (2.1) chybí vysvětlení u S_X^2 a S_Y^2.
- část 2.1.3.: Stejné jako u bodu 1 – místo "hodnoty"náhodných veličin bych spíše používal termín "realizace"náhodných veličin a pro přehlednost značil malými písmeny.
- str.13[3]: u vzorce pro střední hodnotu není vysvětleno, co je to X, takže není dostatečně zřejmé, co má autorka na mysli. Vzhledem k tomu, že na str.13[8] je $p_i = n_i/N$ empirické, tak se celkově spíše jedná o odhad střední hodnoty, než o střední hodnotu jako takovou.
- str.15[7]: Není vysvětleno, co jsou prvky S_i.

Věcné nedostatky:

- Na str.12 v části 2.2.1. bych u popisu proměnných v datasetu přidal odkaz na tabulku 3.1., kde je uveden jejich seznam a vysvětlení
- Úvod kapitoly 3: definice vícero přístupů k segmentaci je k nalezení např. v knize [Jain a Dubes, 1988], na str.56.
- V kapitole 3 bych uvítal alespoň stručný popis nejdůležitější funkce `archetypes` z prostředí `R`, která byla využita v aplikační části k nalezení archetypů.
- V kapitole 3 v aplikační části by se slušelo na str.20 charakterizovat všech 6 nalezených archetypů a ne jenom 3 vybrané, případně popsat rozdíly mezi nimi, a podrobněji charakterizovat jednotlivé segmenty, které vznikly rozdílnými segmentacemi (v práci je uveden pouze rozdíl ve velikosti jednotlivých segmentů).

Závěr
Práci hodnotím jako podprůměrnou. Neměl jsem možnost seznámit se se všemi částmi práce před jejím odevzdáním. Negativně působí vícero nepřesností v kapitole 2 a podle mého názoru ne příliš pečlivý komentář k výsledkům dosažených na reálných datech. Naopak, na celé práci si cením podrobného důkazu v kapitole 1. Dále oceňuji protipříklady, které ukazují nejednoznačnost úlohy archetypální analýzy, a také to, že segmentace, založená na koeficientech α (které vyjadřují procentuální míru příslušnosti k jednotlivým archetypům), není shodná s tím, kdy pozorování zařadíme k archetypu s nejbližší Eukleidovskou vzdáleností. Tyto příspěvky vymyslela a zpracovala autorka zcela samostatně. Datový soubor pro aplikační část rovněž vyžadoval poměrně rozsáhlé předzpracování, které autorka musela provést před samotnou aplikací archetypální analýzy. Doporučuji uznat předloženou práci jako bakalářskou práci ve studijním oboru Obecná Matematika, pokud autorka přinese pro obhajobu errata k předložené práci a podrobněji okomentuje dosažené výsledky v aplikační části.

Marek Dvořák
Katedra pravděpodobnosti a matematické statistiky
2.9.2015