Oponentský posudek na habilitační práci Ing. Zdeňka Šulce , Ph.D.

HIERARCHICAL CLUSTER ANALYSIS OF CATEGORICAL DATA

Oponent: Doc. RNDr. Jaroslav Michálek, CSc., Universita obrany, Brno

Předložená habilitační práce spadá do oblasti výpočetní statistiky, tedy do oboru, jež v posledních dvaceti letech zaznamenal ve světě velmi dynamický rozvoj. Je pro mne potěšující, že mohu hned na začátku svého posudku poznamenat, že Ing. Zdeněk Šulc, Ph.D. se v této oblasti významně podílí na rozvoji klasifikačních metod pro kategoriální data a zejména na jejich dostupnosti zájemcům z řady vědních oborů. Výsledky jeho práce věnované hierarchickému shlukování kategoriálních dat, které obsahují nominální proměnné s více než dvěma kategoriemi, jsou velmi žádoucí v mnoha oborech, zejména v oblasti ekonomických, sociologických, psychologických a lékařských statistických analýz. Ocenění si také zaslouží, že se podílel na návrhu nové míry podobnosti pro kategoriální data a jeho samostatný návrh nového kritéria pro hodnocení shlukování. Zvlášť cenné je, že se autor bezprostředně zabývá také počítačovou implementací popsaných teoretických výsledků a tím umožňuje přímé využití těchto metod zájemcům z aplikačních oblastí. Výsledky práce mohou být rovněž bezprostředně využity při výuce mnohorozměrných statistických metod a jejich aplikací.

Vlastní práce je velmi obsáhlá, obsahuje celkem 129 stran, z toho je 16 úvodních stran, kde lze nalézt abstrakt, obsah, seznamy grafů, tabulek, zkratek a symbolů. Kromě toho je k práci přiložen CD disk, kde lze také nalézt softwarovou aplikaci v práci diskutovaných metod zpracovanou ve výpočetním prostředí R.

Práce je rozdělena do šesti kapitol. První kapitole ještě předchází Úvod, kde autor práce s ohledem na stávající literarní prameny stručně shrnuje problematiku hierarchické shlukovací analýzy pro kategoriální data a s ohledem na neřešené problémy této analýzy formuluje tři základní cíle své dizertace. První je porovnat míry podobnosti používané v analýze shlukování a také srovnání těchto měr s měrami, které ve své práci dříve navrhl autor dizertace. Dále pak doporučení pro výběr vhodné míry podobnosti s ohledem na daný datový soubor, který je třeba podrobit shlukování. Druhým formulovaným cílem, který také vychází z jeho publikovaných výsledků, je srovnání běžně

používaných interních hodnotících kriterií shluků vytvořených z kategoriálních dat a analyzovat jejich vzájemné vztahy. Výsledky porovnání založené na vyhodnocení experimentu provedeného na rozsáhlých generovaných souborech dat pak může pomoci výzkumníkům rozhodnout se v dané situaci pro vhodné hodnotící kritérium. Konečně třetí cíl dizertace je umožnit zájemci popsané a doporučené metody shlukování bezprostředně prakticky používat. K tomu má sloužit balíček programů `nomclust 2.0` vytvořený v prostředí R. Tento nově vytvořený balíček, který byl spolu se spoluautory publikován v prestižním časopise *Computional Statistics* roce 2022, navazuje na jeho dřívější jednodušší verzi a vylepšuje ji.

První dvě kapitoly práce jsou věnovány časopiseckému shrnutí literárních výstupů, které se zabývají shlukováním kategoriálních dat a mírami podobosti zaměřenými na kategoriální data. Diskutovaný přehled časopiseckých výstupů, který autor uvádí, je velmi výstižný, seznam literatury obsahuje 96 položek. Zvlášť elegantně je popsán souhrn v literatuře používaných měr podobnosti pro kategoriální data. Výstižně popisuje a exaktně definuje 16 podobnostních měr pro kategoriální data, uvádí jejich autory, přehledně pomocí tabulky uvádí jejich výpočetní vzorce a jejich vlastnosti s ohledem na charakter dat a na požadované výsledky shlukování. Důležité je poznamenat, že spoluautorem návrhu dvou z těchto měr je autor předkládané habilitační práce. Dále si autor ve druhé kapitole také všímá způsobů vážení kategoriálních proměnných, metod a principů hierarchického spojování shluků. Konečně výsledky srovnání studovaných klasifikačních metod ukazuje také na numerickém příkladě. Obě kapitoly svědčí, že autor práce je s popisovaými metodami a současnými trendy v této oblasti po teoretické i praktické stránce velmi dobře obeznámen, má dobrý přehled časopiseckých zdrojů a vhodně uplatňuje svůj kritický nadhled.

Třetí kapitola je věnována kriteriím pro vyhodnocování shluků. Na ilustrativním příkladě popisuje dvě externí hodnotící kritéria, která jsou v praxi často využívána. Dále se detailně zabývá popisem interních hodnotících kritérií, která vycházejí z podobnosti nebo naopak z odlišností objektů v daném shluku. Přehledně uvádí 11 kritérií, uvádí jejich zavedení, vzorce a také jejich interpretaci s ohledem na stanovení počtu shluků. Zvlášť cenné je, že v práci dále navrhuje také dvě vlastní kritéria, která se ukazují jako vhodná pro stanovení počtu shluků.

Ve čtvrté kapitole je popsán výpočetní balíček programů `nomclust 2.0`, který představuje novou generaci dřívě vytvořeného balíčku programů označeného `nomclust 1`. Balíček je věnován komplexní analýze shlukování kategoriálních dat, detailně jsou popsány jednotlivé obsažené funkce a způsob jejich volání. Jsou k dispozici výpočty měr podobnosti z kapitoly 2 a také možnost výpočtu třinácti hodnotících kritérií (externích i interních) popsaných v kapitole 3. Také je k dispozici grafický výstup vybraných procedur, např. dendrogram pro zvolenou míru podobnosti a zvolený počet shluků. Konečně v rámci experimentu, který byl proveden na 60 datových souborech, byl měřen výpočetní čas při

použití předchozího balíčku programů `nomclust 1.0` a při použití nově vytvořeného balíčku programů `nomclust 2.0`. V rámci tohoto experimentu bylo ukázáno, že výpočetní čas jednotlivých měr podobnosti při daných technických parametrech počítače se při použití `nomclust 2.0` proti použití `nomclust 1.0` zkrátil více než stokrát. Možnost bezprostředního použití tohoto balíčku je zvlášť užitečná pro praktické uživatele klasifikačních metod. V práci je tento balíček také využit ke srovnání výše popsaných podobnostních měr pro kategoriální data na 2 700 generovaných datových souborech. Toto srovnání je uvedeno v kapitole 5, je zde popsán proces generování dat, dále je zde uvedena metodologie pro hodnocení měr podobnosti a konečně je popsán vlastní experiment a jeho výsledky s praktickými doporučeními, ve kterých situacích jsou jednotlivé míry podobnosti vhodné a ve kterých ne s ohledem na vybranou hierarchickou metodu spojování shluků. V závěru této kapitoly autor formou tabulek popisuje, které spojení podobnostní míry s typem shlukovací metody při vyhodnocovacím indexu PSFE založeném na entropii a následně založeném na indexu utility CU vychází nejlépe a dále uvádí uspořádané hodnocení kvality jednotlivých kombinací podobnostní míry a metody shlukování (tedy celkem srovnává 48 těchto kombinací).

V závěrečné šesté kapitole se věnuje srovnání hodnotících kritérií, tedy problematice, kterou dříve zformuloval jako cíl 2. Srovnání vychází z porovnání výsledků získaných na 8 100 generovaných datových souborech. Výsledky získaných výstupů hierarchické shlukové analýzy potom porovnával pomocí korelací a metodami analýzy rozptylu. Získané závěry, které jsou užitečné při praktickém provádění shlukovací analýzy, přehledně uvádí v závěru této kapitoly. S ohledem na požadavky uživatele (kvalita shluků, optimální počet shluků) uvádí doporučená hodnotící kritéria.

Celá práce je psána velmi přehledně, s dostatečným nadhledem a svědčí, že autor hluboce pronikl jak do statistické tak do výpočetní oblasti studované problematiky. Tři cíle stanovené v úvodu práce byly s nadhledem splněny. Výsledky uvedené v kapitolách 2 a 5 představují splnění cíle 1, podobně výsledky uvedené v kapitole 3 a kapitole 6 ukazují, že byl také plně splněn cíl 2 a konečně cíl 3 byl splněn vytvořením balíčku nové generace programů `nomclust 2.0`, který je popsán v kapitole 4 a je k dispozici je na přiloženém CD disku. Na práci zejména oceňuji, že autor kvalitně a systematicky popsal metodologii hierarchického shlukování kategoriálních dat, která se v takto ucelené formě v literárních pramenech takto systematicky nevyskytuje. Jako spoluautor se podílí na zavedení dalších dvou měr podobnosti pro kategoriální data a samostatně navrhl dvě nová kritéria pro stanovení počtu shluků. Navíc zájemcům o hierarchickou analýzu kategoriálních dat poskytnul souhrnné teoretické zázemí a bezprostřední balíček programů pro praktické využití uvených metod. Předpokládám, že tento balíček zvláště ocení uživatelé z oblasti ekonomie a zájemci, kteří vyhodnocují medicínská, sociologická a psychologická data. Velmi cenným výsledkem předložené práce je také její bezprostřední

využitelnost v pedagogické činnosti. Studenti se mohou na jednom místě seznámit s postupy užívanými při shlukování kategoriálních dat a pomocí předloženého balíčku programů si mohou popsané postupy na svých datech jednoduše ověřit.

Závěr: Vzhledem k odborné kvalitě výsledků uvedených v práci a celkovému odbornému profilu autora doporučuji práci k obhajobě před Vědeckou radou a jednoznačně doporučuji, aby předložená práce byla přijata za habilitační a aby Ing. Zdeněku Šulcovi, Ph.D. byla na jejím základě udělena hodnost docenta pro obor Statistika.

Jaroslav Michálek V Brně 30. 1. 2024