V ý p i s

ze zápisu ze zasedání vědecké rady Fakulty informatiky a statistiky VŠE v Praze,
konané dne 4. 4. 2024

Přítomni: dle prezenční listiny

Program: ad 2) Habilitační řízení Ing. Zdeňka Šulce, Ph.D., docentem pro obor Statistika

ad 2. Habilitační řízení Ing. Zdeňka Šulce, Ph.D. – obor „Statistika“

Děkan Fakulty informatiky a statistiky prof. Ing. Jakub Fischer, Ph.D. seznámil členy vědecké rady se složením habilitační komise, která posuzovala materiály k habilitačnímu řízení Ing. Zdeňka Šulce, Ph.D. Předsedkyní komise byl jmenována prof. Ing. Stanislava Hronová, CSc., členy komise byli prof. RNDr. Marie Demlová, CSc., prof. Ing. Jana Hančlová, CSc., doc. Ing. Mgr. Jirka Janová, Ph.D. a prof. Mgr. Erik Šoltés, PhD. Prof. Hronová, prof. Hančlová a prof. Šoltés byli osobně přítomni jednání vědecké rady, prof. Demlová a doc. Janová se z jednání řádně omluvili. Osobně byli přítomni i dva oponenti, a to doc. RNDr. Jaroslav Michálek a doc. Ing. Mária Vojtková, PhD.

Poté děkan předal slovo prof. Hronové, která představila habilitanta a přednesla výtah ze zprávy habilitační komise. Kritéria jsou splněna s velkou rezervou jak v oblasti vědecké, tak v oblasti pedagogické. Ing. Šulc publikoval 11 IF článků (některé ve špičkových časopisech) plus dalších 5 v časopisech ve Scopusu; dosáhl 166/172 citací (WoS/Scopus); pracoval ve dvou grantech GAČR; vedl výuku jak v bakalářském, tak v magisterském stupni, česky i anglicky. Je garantem pěti předmětů v anglickém jazyce vyučovaných Katedrou statistiky a pravděpodobnosti. Dále je autorem jedněch skript a balíčku pro programovací jazyk R. V letech 2014, 2018 a 2022 se účastnil třech vědeckých pobytů v zahraničí (jednou na Matematickém institutu Univerzity Leiden a dvakrát na Institutu Psychologie Univerzity Leiden). Je členem České statistické společnosti a redakční rady časopisu Acta Informatica Pragensia. V roce 2017 získal cenu děkana za nejlepší disertační práci a v roce 2019 cenu děkana za nejlepší recenzovaný článek. V roce 2022 se stal finalistou soutěže Cluster Benchmarking Challenge, díky čemuž byl vyzván prezentovat svůj výzkum na konferenci International Federation of Classification Societies (IFCS) 2022. Dále prof. Hronová představila habilitační práci – *Hierarchical Cluster Analysis of Categorical Data* a sdělila, že všechny tři posudky jsou kladné. Dále uvedla, že komise zhodnotila práci uchazeče a v tajném hlasování se jednomyslně usnesla (všemi pěti kladnými hlasy), že jmenování Ing. Šulce docentem podporuje.

Ing. Šulc přednesl habilitační přednášku s názvem *Shluková analýza kategoriálních dat*, ve které představil vybrané algoritmy pro shlukování kategoriálních dat. U každého algoritmu uvedl jeho princip a dále jeho silné a slabé stránky. V závěru prezentace shrnul hlavní vlastnosti představených algoritmů a doporučil, pro jaké situace jsou jednotlivé algoritmy vhodné.

V následné diskusi vystoupili:

- *prof. Gregorová: Existuje alternativa k metodě k-modů, podobně, jako metoda k-means++ k metodě k-means?* Ing. Šulc uvedl, že metodu k-means++ nezná. Následně jmenoval několik alternativních metod k metodě k-modů a vysvětlil jejich princip.
- *prof. Jablonský – Na jakém principu funguje určování optimálního počtu shluků?* Ing. Šulc uvedl, že optimální počet shluků se určuje pomocí kritéria BIC v případě metody LCA založené na modelu, nebo pomocí jeho modifikace kritéria BIC založené na variabilitě vzniklých shluků, která se používá u dvoukrokové shlukové analýzy.

- *prof. Černý – Na jakém principu je založena metoda LCA?* Ing. Šulc vysvětlil obecně princip metody LCA. Následně chtěl prof. Černý znát detaily této analýzy, např. její předpoklady. Ing. Šulc odpověděl, že se při shlukování běžně žádné předpoklady neověřují.
- *prof. Svátek – Je možné představené shlukové algoritmy použít v situaci, pokud mají proměnné hierarchickou strukturu kategorií?* Ing. Šulc odpověděl, že hierarchickou strukturu kategorií nezná. Po upřesnění otázky ze strany prof. Svátka Ing. Šulc navrhl několik postupů, jak by se taková data dala shlukovat. Tyto návrhy byly ale prof. Svátkem rozporovány.
- *prof. Cipra – Proč existuje pouze několik shlukových algoritmů, které doporučují optimální počet shluků?* Ing. Šulc uvedl, že optimální počet shluků je pouze doporučení pro výzkumníka, s jakým počtem shluků má začít při shlukové analýze. Dále uvedl několik postupů, jak může výzkumník určit optimální počet shluků v případě, že daná metoda optimální počet shluků nedoporučuje.
- *prof. Hynek – Jsou rychlé a rychlejší algoritmy shlukové analýzy, lze uvést ten nejrychlejší?* Ing. Šulc zhodnotil rychlost jednotlivých představených algoritmů a následně jmenoval dva nejrychlejší algoritmy, které jsou založeny na agregaci údajů o shlukovaných datech.

Následovala přednáška, v níž Ing. Šulc představil svoji habilitační práci. Nejprve vysvětlil, že se v rámci své práce věnoval aglomerativnímu hierarchickému shlukování kategoriálních dat. Poté představil tři hlavní cíle své práce, a to analýzu v oblasti měr podobnosti pro kategoriální data, analýzu interních hodnoticích kritérií pro kategoriální data a představení a další zdokonalení balíčku pro programovací jazyk R, který se zabývá komplexní analýzou kategoriálních dat. Dále představil výzkum, ze kterého vycházel, a prezentoval vlastní články, které tento výzkum rozvíjely. V další části přednášky představil vybrané výsledky, kterých bylo v práci dosaženo. V závěru shrnul hlavní přínosy práce.

Děkan prof. Fischer poté předal slovo přítomné oponentce doc. Vojtkové. Doc. Vojtková konstatovala, že předložená práce splňuje požadavky kladené na habilitační práce z hlediska obsahu, formy, přínosu a stanoveného cíle. Předloženou práci doporučila k obhajobě. V posudku měla připravené dvě otázky, ale protože první z nich považovala prostřednictvím habilitační přednášky za zodpovězenou, položila pouze druhou otázku: *Jaký je Váš názor na použití programovacího jazyka Python resp. některého z komerčních balíkových programů jako např. SPSS nebo SAS při aplikaci těchto metod?* Ing. Šulc odpověděl, že všechny použité metody v habilitační práci je technicky možné naprogramovat v Pythonu. V současné době ale tato implementace chybí, takže by bylo potřeba původní kód v R přepsat do Pythonu a mírně jej upravit. Software SPSS neobsahuje zkoumané míry podobnosti ani většinu hodnoticích kritérií. Standardně jej tak nelze použít. Pro použití těchto metod by bylo potřeba vypočítat odděleně matici nepodobností, např. v R nebo Pythonu, a následně ji použít jako vstup pro hierarchické shlukování v SPSS.

Následně děkan prof. Fischer předal slovo přítomnému oponentovi doc. Michálkovi. Ten zejména vyzdvihl, že autor v habilitační práci kvalitně a systematicky popsal metodologii hierarchického shlukování kategoriálních dat, která se v takto ucelené formě v literárních pramenech nevyskytuje. Dále ocenil balíček pro programovací jazyk R, který mohou využít zájemci z různých oborů a také studenti. Vzhledem k odborné kvalitě výsledků uvedených v práci a celkovému odbornému profilu autora práci doporučil k obhajobě. Následně položil otázku: *Máte zkušenosti s mírami podobnosti pro smíšená data obsahující kvantitativní i kategoriální proměnné? Jak se tyto míry chovají při různých poměrech kvantitativních*

a kategoriálních proměnných? Ing. Šulc odpověděl, že měr podobnosti pro smíšená data neexistuje mnoho. Zmínil Gowerovův koeficient, který pro kategoriální data používá známý princip koeficientu prosté shody. Následně uvedl, že v předchozím výzkumu představil několik modifikací Gowerova koeficientu, u kterých zjišťoval, jak se chovají při různých poměrech kategoriálních a kvantitativních proměnných. Některé z těchto modifikací překonávaly původní verzi Gowerova koeficientu.

Posudek prof. Debické komentovala předsedkyně komise prof. Hronová. Velmi detailní posudek popisuje splnění hlavních třech cílů práce s konstatováním, že všechny cíle práce byly splněny. V rámci druhého cíle, týkajícího se analýzy interních hodnoticích kritérií, nicméně konstatuje, že výběr použitých metod pro analýzu hodnoticích kritérií by měl být lépe odůvodněn. V závěru shrnuje, že habilitační práce obsahuje originální výzkum, který rozšiřuje nástroje pro shlukování kategoriálních dat. Dále uvádí, že práce splňuje požadavky kladené na habilitační práci a že ji doporučuje k obhajobě. V posudku položila tři otázky: *Mohly by být míry podobnosti citlivé na rozdělení četností kategorií klasifikovány na základě hodnoty indexu struktury nerovnosti?* Ing. Šulc odpověděl, že tento konkrétní index prozatím ve svém výzkumu nepoužil. Zabýval se ale velmi podobným konceptem, kdy zkoumal závislost měr podobnosti na počáteční variabilitě v datovém souboru. Výsledky ale ukázaly, že zkoumané míry podobnosti nezávisejí na počáteční variabilitě v souboru, a nemohou tak být podle ní klasifikovány. *Jaké další míry závislosti by mohly být použity pro hodnocení hodnoticích kritérií v kapitole 6.2?* Ing. Šulc odpověděl, že v případě poměru determinace (eta-squared) by mohly být použity míry založené na podobném principu, např. omega-squared. Namísto Pearsonova korelačního koeficientu lze použít neparametrický Spearmanův korelační koeficient. *V jakých případech by habilitant preferoval použití kritérií HE a HM namísto interních kritérií představených v sekci 3.2.1.?* Ing. Šulc odpověděl, že kritéria HE a HM mohou být použita při hledání optimálního počtu shluků v datovém souboru. Zmínil ale, že existují lepší alternativy než zmíněná kritéria, např. kritéria BIC nebo BK, která by měl výzkumník vyzkoušet přednostně.

Následovala diskuse.

- *prof. Svátek – Článek, který jste zmiňoval ve Vaší prezentaci a který je mimo poklady k habilitačnímu řízení, byl skutečně přijatý?* Ing. Šulc odpověděl, že uvedený článek byl skutečně přijat k publikaci.
- *prof. Gregorová – ocenila kvalitní přednášku. Měla několik dotazů:*

  Ve své přednášce jste zmínil, že míry pro binární data jsou velmi podobné nebo dokonce stejné. Byly tyto závěry zjištěny empiricky nebo také teoreticky? Ing. Šulc odpověděl, že u některých měr podobnosti pro binární data byla dokázána shoda teoreticky. Většina těchto měr však není úplně shodná, ale jejich hodnoty jsou velmi silně korelovány. To často vede k situaci, že shluky získané prostřednictvím těchto měr jsou totožné.

  Jaký byl počet objektů v syntetických datových souborech? S jakým největším datovým souborem, co se týče počtu pozorování a proměnných, jste pracoval? Je zde nějaký technický limit? Ing. Šulc odpověděl, že syntetické soubory byly generovány s 600 objekty. Dále odpověděl, že obvykle pracuje s typickými datovými soubory s ne příliš vysokým počtem proměnných a s maximálně deseti kategoriemi. Technicky zde ale není žádné omezení, které by omezovalo velikost datového souboru, počet kategoriálních proměnných nebo jejich hodnot.

  Nejsou zvýhodněny takové míry podobnosti, které byly použity pro generování syntetických datových souborů? Ing. Šulc odpověděl, že syntetické datové soubory byly původně generovány jako kvantitativní a byly až následně kategorizovány. Proto

nehrozí, že by v provedených experimentech byla některá ze zkoumaných měr podobnosti zvýhodněna.

Existuje v balíčku R nomclust implementace, která automaticky výzkumníkovi doporučí např. vhodnou míru podobnosti? Ing. Šulc odpověděl, že tato implementace není do balíčku zakomponována, ale že jsou ve výchozím nastavení funkcí v balíčku přednastaveny takové míry podobnosti, které obvykle vedou k dobrým výsledkům.

Po ukončení této části proběhla neveřejná diskuse.

Po ukončení této části vyzval děkan fakulty v neveřejné části zasedání členy vědecké rady k tajnému hlasování, jehož výsledek je následující:

- počet členů vědecké rady: **44**
- počet členů VR přítomných: **36**
- počet odevzdaných hlasů: **36** kladných
  0 neplatných
  0 záporných

USNESENÍ: Vědecká rada FIS schvaluje návrh na jmenování Ing. Zdeňka Šulce, Ph.D. docentem pro obor Statistika.

Děkan Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze předloží podle § 72 odst. 11, zákona č. 111/1998 Sb. rektorovi Vysoké školy ekonomické v Praze návrh na jmenování

Ing. Zdeňka ŠULCE, Ph.D.
d o c e n t e m
pro obor Statistika