Oponentský posudek na diplomovou práci

Název: Klasifikace na základě longitudinálních pozorování
Autor: Lukáš Bandas

Shrnutí:

Diplomová práce Lukáše Bandase se zabývá metodami diskriminační analýzy pro longitudinální data. Po krákém motivačním úvodu následuje kapitola zabývající se stručným popisem lineáního smíšeného modelu a metodami pro odhadování jeho parametrů. Další kapitola se týká diskriminační analýzy. Nejprve jsou na několika stranách shrnuty koncepty klasické diskriminační analýzy pro nezávislé a stejně rozdělené vektory. Následuje vlastní jádro práce: popis čtyř přístupů k diskriminační analýze longitudinálních dat v kapitole 3.2. V kapitole 4 jsou tyto metody aplikovány na reálná data. Práci uzavírá krátká kapitola shrnující výsledky simulačních studií a závěrečné poznámky.

Práce má popisný charakter. Přehledně shrnuje vybrané metody pro diskriminační analýzu longitudinálních dat a ukazuje jejich použití v praxi. Po formální stránce je většinou velmi pěkná, zejména její první polovina je zpracována celkem pečlivě. Překlepy a gramatické chyby se najdou, ale není jich nepřiměřeně mnoho.

Chybí mi však hlubší vhled do studované problematiky a snaha o celkovou syntézu. Celá práce zůstává na povrchu a nesnaží se o hlubší pochopení. Její jednotlivé části mají izolovaný charakter a nejsou dostatečně propojeny. Řada důležitých otázek je odbyta krátkými zmínkami nebo je zcela ignorována. Aplikace na data je provedena naprosto mechanisticky bez uvažování o tom, co se dělá a proč. V simulační kapitole není vůbec popsáno, jak a za jakých předpokladů byla simulovaná data generována. Simulační studie je vůbec natolik odbytá, že by snad bylo lepší, aby pátá kapitola v práci vůbec nebyla. Jelikož Lukáš Bandas se nesnaží o vlastní příspěvek formou rozšíření nebo vylepšení metod pro diskriminační analýzu longitudinálních dat, tím spíše zamrzí, že se v simulační studii nepokusil o důkladnější srovnání popisovaných metod v různých situacích a o studii jejich chování při nesplněných předpokladech (ne každý spojitý vektor je totiž normálně rozdělený).

Svoje obecné námitky bych ozřejmil následujícími zásadními připomínkami:

- V kapitole 2.2 není nic o testování hypotéz a submodelů, a strategii volby modelu. Volba vhodného modelu má však zásadní význam pro klasifikaci. V kapitole 4 jsou uvedeny konkrétní modely, ale není vysvětleno, jak byly získány a jak byly ověřeny jejich předpoklady.
- V kapitole 3.2 (ani později) se neuvažuje počet a rozmístění časů pozorování. Kdyby se klasifikační skupiny v trénovacích datech systematicky lišily počtem pozorování a rozmístěním časů, klasifikace by asi moc dobře nefungovala.
- V poznámce 3.8 na str. 20 je zmíněno, že předpoklad nezávislosti residuí v longitudinálních datech není nutný a je nevhodný. S tím lze jen souhlasit. Proč se tedy s tímto předpokladem pracuje a proč se klade i na bilirubin a albumin v aplikacích? Byl tento předpoklad ověřován?
- Kapitola 3.2.4 (funcionální přístup) je zcela izolovaná od zbytku práce. Na několika místech se odvolává na „analogii“ s předchozí kapitolou, ale nikdy není detailně vysvětleno, v čem ta analogie spočívá. Metody z kapitoly 3.2.4 se nikdy neobjevují v aplikacích ani v simulacích a ve zbytku práce se o nich nehovoří.
- Aplikace na praktický příklad klasifikace lidí na „živé“ a „mrtvé“ je velmi problematická. Zatímco celý teoretický přístup spoléhá na to, že pozorování lze „a-priori“ rozdělit do dvou skupin a přiřazení do skupiny je pevné, nyní se celá metodika aplikuje na situaci, kdy rozdělení do skupin je časově podmíněno („živý“ se může stát „mrtvým“) a vše je navíc komplikováno cenzorováním transplantacemi jater, které nejspíše závisí jak na předchozích hodnotách jaterních

testů tak na riziku úmrtí. Nekladenou a nezodpovězenou otázkou je, jak potom takovouhle klasifikaci vůbec interpretovat.

Diplomová práce Lukáše Bandase minimalistickým způsobem naplňuje zadání, celkově ji hodnotím jako uspokojivou a doporučuji ji uznat jako práci diplomovou.

Drobné připomínky:

- 5_{11} *„jsou použity data“*
- 11_1 REML věrohodnost se maximalizuje pro pevné $\widehat{\boldsymbol{\beta}}$ nebo se bere v úvahu, že $\widehat{\boldsymbol{\beta}}$ je funcí $\boldsymbol{\alpha}$?
- Kapitola 3.1: nebere se v úvahu, že hustotu pozorování neznáme. Chybí vysvětlení účelu trénovacího souboru.
- 20, (3.9): co je $\boldsymbol{t}_i$?
- 22, (3.13): obě + měly být −.
- 22, pod (3.15): překlepy v $\boldsymbol{T}$ a (.
- 23_1 Původně bylo λ, teď najednou $\lambda_{i1}, \lambda_{i2}, \ldots$?
- 24^2 Musí existovat taková posloupnost ortonormálních funkcí?
- 24_4 *„Pokud nás zajímá analogie s lineárním smíšeným modelem, tak koeficienty ξ_{il} hrají roli náhodných efektů $\boldsymbol{b}_i$ a vlastní funkce ϕ_{il} hrají roli matice $\mathbb{Z}_i$.“* Dá se to ukázat?
- 25, (3.24): Z čeho plyne (3.24) a co znamená Y_i?
- 25, (3.25): Z čeho plyne (3.25)?
- 26^5 *„rozdíly mezi skupinami subjektů v rámci disjunktních skupin“* Nesmyslná věta.
- 27_{11} *„Přitom se zaměříme pouze na pacienty, kteří přežili bez transplantace 910 dnů.“* Proč zrovna 910 dnů?
- 28^5 *„Kvalita klasifikace byla ohodnocena metodou cross-validation“* Chybí vysvětlení: co to je, a jak to bylo provedeno?
- 29 Zdá se, že u mrtvých občas nejsou k dispozici data až do konce, zatímco u živých ano. Není to problém?
- 31_{16} *„V Tabulce 4.2 můžeme sledovat chybovost zařazování, senzitivitu a specificitu.“* Co se pod těmito pojmy míní?
- 31_9 *„. . . u všech metod vychází poměrně vysoká specificita, což nám vlastně říká, že nízké procento pacientů, kteří jsou pozitivní (tedy v našem případě nepřežijí dané období osmi let, tj. skupina 2), označíme jako negativní, tedy zařadíme do skupiny 1.“* Není to naopak?
- 32, tabulka 4.1: V tabulce chybí odhadnuté hodnoty var $\boldsymbol{b}_i$.
- 39 Proč je v modelu (4.4) kvadratický člen? Podle tabulky 4.7 není významný.
- 47 Co znamená sloupec *„error“*?

doc. Mgr. Michal Kulich, PhD.
KPMS MFF UK
6. května 2012