Posudek diplomové práce

Jaroslav Ševčík

Směsové modely pravděpodobnostních rozdělení.

V diplomové práci se autor zabývá směsovými modely, které jsou používány nejčastěji k modelování populací skládajících se z více „subpopulací". Předpokládaný model pak zahrnuje určitá podmíněná rozdělení, pravděpodobnostní rozdělení sledované náhodné veličiny za podmínky, že patří k dané subpopulaci a marginální rozdělení pravděpodobností jednotlivých subpopulací. Hlavním problémem zde bývá odhad počtu složek modelu (subpopulací) a poté odhad jednotlivých parametrů zmíněných podmíněných rozdělení. Netriviální je i výpočetní stránka věci.

Diplomant rozdělil práci do tří kapitol. V první je připomenuta teoretická stránka směsových modelů, jejich vlastnosti a možná použití. Důležitou součástí je pojednání o odhadech parametrů směsových modelů za předpokladu známého počtu složek směsi. Ve druhé kapitole se řeší zásadní problém současného odhadu počtu složek a parametrů modelu. Je popsána celá řada metod a diskutována i jejich výpočetní náročnost. Ve třetí části jsou praktické ukázky použití směsí na simulovaných i reálných datech, autor představuje i některé své vlastní programy v prostředí R.

Práce mě překvapila počtem popisovaných metod. Diplomant musel nastudovat nesmírné množství literatury, aby roztřídil jednotlivé přístupy ke směsovým modelům. Na druhou stranu právě množství popisovaných metod vede k tomu, že nezbývá prostor pro hlubší analýzu jednotlivých přístupů a jejich porovnání. Možná by stačilo popsat různé způsoby analýzy směsových dat a v rámci takových obecných přístupů pak detailněji studovat jeden až dva konkrétní algoritmy. Nicméně v tomto bodě se jedná o můj osobní názor a nepovažuji to za závažný nedostatek práce.

Více mi vadí, že práce obsahuje některé vágní formulace. Například na straně 4 lze číst: „Predpokladajme, že hustotu $f(x)$ môžeme zapísať v tvare $f(x) = \sum \pi_i f_i(x)$, kde f_i sú hustoty a π_i nezáporné čísla so súčtom 1". Problém je v tom, že až na triviální případ degenerovaného rozdělení **každou** hustotu můžeme napsat v takovém tvaru (a to odhlédnu-li od faktu, že není předpokládáno, že f_i jsou různé) a uvedený rozklad až na výjimky není jednoznačný, ani počet možných komponentů směsi není pevný. V práci se sice později diskutuje pouze případ, kdy jednotlivé hustoty f_i patří do nějaké parametrické rodiny, nemůže tedy jít o libovolná rozdělení, ovšem problém jednoznačnosti rozkladu hustoty f na konvexní kombinaci prvků povolené parametrické třídy není ani potom diskutován.

Překvapuje mě, že při množství popisovaných postupů jsem nenašel důkladnou diskusi o přínosu tohoto parametrického přístupu proti neparametrickým odhadům hustoty, zejména v interpretaci výsledků. Jde o můj osobní názor, ale parametrické směsové modely jsou velmi ovlivněné volbou typu rozdělení komponent, což snižuje mou důvěru k nim.

Některé další poznámky:

- Na straně 15 se hovoří o tom, že t-rozdělení je robustní alternativou normálního rozdělení a na straně 33 jsou stupně volnosti pro t-rozdělení označeny za parametr robustnosti. Oboje vyžaduje vysvětlení, protože slovo robustní má dost specifický význam.
- Práce obsahuje dost překlepů, které mohly snadno být odstraněny kontrolou pravopisu.
- Na straně 39 se hovoří o konvergenci bodu k množině. Korektnější je definovat vzdálenost bodu od množiny a tuto konvergenci převést na konvergenci vzdálenosti bodu od množiny nule.
- Na straně 59 se zavádí předpoklad (P): „spĺňa nieľko ďalších podmienok" s poznámkou pod čarou, že podmínky jsou analytického charakteru a komplikované na ověření. Víme alespoň, které třídy rozdělení tyto podmínky splňují?

Celková úroveň práce je vysoká, jedná se o velmi kompletní a rozsáhlou studii velkého množství metod a z tohoto hlediska považuji práci za velmi přínosnou. Protože práce splňuje požadavky kladené na diplomovou práci, doporučuji ji přijmout k obhajobě.

Daniel Hlubinka
22. září 2006