SOUČASNÝ STAV ROBUSTNÍCH PROGRAMŮ PRO GNOSTICKOU ANALÝZU DAT

P.Kovanic, ÚTIA ČSAV, Praha

1. Úvod

Gnostická teorie dat byla vytvořena jako alternativa matematické statistiky pro datové soubory, které mají neznámý statistický model nebo které ani nemohou být dobře charakterizovány statistickým modelem. Z této teorie byla již odvozena řada algoritmů, ověřených v různých aplikacích na simulovaných i na reálných datech včetně porovnání se statistickými algoritmy. Účelem tohoto příspěvku je podat stručný přehled o stavu prací na gnostických programech.

2. Stručný přehled gnostických programů

2.1. Základní gnostické procedury

Základními veličinami, které je třeba při gnostické analýze dat odhadovat, jsou parametr polohy a měřítka datového souboru. Tyto veličiny jsou v gnostické teorii definovány poněkud jinak, než ve statistické teorii.

Datový soubor je n-tice dat z_i $(i=1,\ldots,n)$ mající model

$$z_i = z_0 \exp(s\Omega_i) \qquad z_0 \in R_+,\ s \in R_+,\ \Omega_i \in R_1$$

kde z_0 je parametr polohy a veličina s je parametr měřítka.

Z gnostické teorie rovněž vyplývá vzorec pro distribuční funkci datového souboru, která je analogií odhadu pravděpodobnostní distribuce. Gnostická distribuční funkce závisí kromě dat jen na parametru měřítka. Může být proto použita k odhadování parametru měřítka řešením rovnice, získané z podmínky nejlepší shody gnostické a empirické distribuční funkce

datového souboru. Procedury pro výpočet gnostické distribuční funkce a pro odhadování parametru měřítka na uvedeném principu jsou pak základem všech gnostických programů.

Pro odhadování parametru polohy byla vyvinuta řada procedur využívajících gnostické veličiny. Všechny potřebují parametr měřítka, který může být buď odhadnut z dat nebo zadán. Druhá z těchto možností se využívá při analýze shluků dat, jakož i v případech, kdy je odhad parametru měřítka znám z předchozích analýz. Gnostické procedury pro odhad [illegible] polohy určují parametr polohy jako polohu maxima hustoty dat, která je derivací gnostické distribuční funkce dat. Řešení této úlohy může být podmíněno splněním požadavku unimodality hustoty, nebo se hledá nepodmíněně. Obě verze jsou k dispozici. Nepodmíněná maximalizace se používá zejména v testech unimodality. Stupeň robustnosti vůči odlehlým pozorováním lze u některých procedur zadat.

Čtvrtou ze základních gnostických procedur je algoritmus hustoty dat. Ten má dvě verze, první počítá hodnoty derivace průměru distribučních funkcí datového souboru, druhá derivaci distribuční funkce celého datového souboru, tj. funkce získané aplikací kompozičního axiomu gnostické teorie. První z těchto verzí je vhodnější pro analýzu souborů obsahujících více shluků, druhá slouží k odhadování hustoty homogenních souborů s unimodální hustotou.

2.2. Sekundární gnostické procedury

Tyto procedury využívají základní procedury k řešení dalších úloh. Mezi ně patří zejména:

- testy homogenity datového souboru
- odhady mezí pásma typických hodnot dat a klasifikace dat na typická, podtypická a nadtypická
- testy příslušnosti dat k datovému souboru
- testy shody datových souborů

Všechny tyto úlohy se řeší s využitím podstatné nelinearity a robustnosti gnostických odhadů, jde tedy o jiné postupy, než jakých využívá matematická statistika.

2.3. Pomocné procedury

Pro práci s gnostickými procedurami potřebujeme ještě pomocné procedury, zejména pak podprogramy pro

- zavedení dat do počítače a manipulace s daty (vytváření pracovních souborů, jejich modifikace a doplňování a pod.)
- uspořádání dat
- exponenciální transformace aditivních dat na multiplikativní
- logaritmická transformace multiplikativních dat na aditivní
- zobrazování výsledků v grafické i tabelární formě
- řízení operací

2.4. Gnostické programové systémy

Ze základních, sekundárních a pomocných procedur bylo již sestaveno několik gnostických programových systémů pro různé aplikace.

2.4.1. Program GAD pro gnostickou analýzu dat

Tento program je koncipován jako interaktivní programový systém pro hloubkovou analýzu konkrétních datových souborů s použitím osobního počítače s operační pamětí 48 K. Může pracovat jak s aditivně, tak i s multiplikativně kontaminovanými daty a zahrnuje všechny již uvedené základní sekundární i pomocné procedury. Je používán jednak jako pracovní program pro zkušební aplikace a jednak jako prostředí pro ladění a ověřování nových gnostických procedur. Pomocí tohoto programu byly získány dobré zkušenosti při zpracování reálných dat z různých oborů (biochemie, fyziologie rostlin, energetika, scientometrie, chemická technologie aj.).

2.4.2. Gnostické monitory procesů

Tímto názvem se označují programové systémy určené ke zpracování časových řad silně kontaminovaných dat, plnících tyto funkce:

- robustní filtrace dat
- diagnostika měřicího systému
- diagnostika stavů monitorovaného objektu

K plnění těchto funkcí využívá gnostický monitor programy pro robustní odhadování parametrů

měřítka, polohy, distribuční funkce dat a pro odhady mezí pásma typických hodnot dat. Významná je kromě robustnosti všech funkcí i plná adaptivita ke změnám úrovně monitorovaného procesu i ke změnám variability poruch dat. Zatím byly vyvinuty dvě verze:GM1 a GM2. První pracuje s n-ticí nejnovějších dat braných s plnou vahou, druhý s exponenciálně zapomínanými daty. Program GM1 pracuje spolehlivě i při opakovaných výpadcích měřicího systému, avšak je pomalejší. Program GM2 je rekurzivní a pracuje podstatně rychleji. Program GM1 byl předán k obchodnímu využití bratislavskému Datasystému. Mezi nesporně významné aplikace systému GM1 bude patřit využití v řídicím a havarijním systému transitního plynovodu.

2.4.3. Gnostická identifikace regresního modelu

Tato úloha byla úspěšně vyřešena teoreticky pro jakýkoliv (i nelineární) spojitý a diferencovatelný model s vektorovou nezávislou (vysvětlující) proměnnou a skalární závislou proměnnou. Pouze dva prvky řešení odlišovaly gnostický postup od standardního:

1) Chyba aproximace změřených dat závisle proměnné modelem se určuje nikoliv v euklidovské metrice, nýbrž v metrice vyplývající z gnostické teorie (chybou je estimační irrelevance).

2) Kriteriální funkce hodnotící shodu teoretických hodnot s experimentálními je rovněž gnostického typu místo nejběžnějšího součtu čtverců odchylek. (Použije se součet věrností nebo jejich čtverců, nebo součet informačních ztrát).

Řešení takto zformulované úlohy je značně robustní vzhledem k silným poruchám jak závisle, tak nezávisle proměnné. K demonstraci kvality tohoto přístupu byly vypracovány dva programy GI0 a GI1 pro identifikaci lineárního modelu nulového a prvního řádu (bez konstantního členu a s ním) za silných poruch pozorování závisle i nezávisle proměnné.

2.4.4. Robustní systém automatické regulace

Zpětnovazební automatická regulace se běžně realizuje s využitím regulátoru operujícího na regulační chybě. Akční veličina korigující stav řízeného objektu je zpravidla lineární funkcí chyby regulace, přičemž tato funkce se vybírá z podmínky kvality regulace. Chyba regulace je přitom určována stejně, jako orientovaná vzdálenost v euklidovské geometrii, tj. jako rozdíl žádané a skutečné hodnoty. Pracuje-li takový systém v podmínkách silných poruch vstupních, výstupních i pozorovaných proměnných, přenáší se tyto poruchy do regulované proměnné a kvalita regulace se snižuje. Použije-li se však k řízení regulační chyba určená v gnostické metrice,zvýší se podstatně robustnost systému automatické regulace k náhodným poruchám. K demonstraci této vlastnosti byl vypracován program GR1 simulující ekologický systém zajíci-lišky a jeho automatické řízení pomocí nelineární zpětné vazby. Tento systém je popsán diferenčními rovnicemi Volterra-Lotkova typu, jehož koeficienty jsou zčásti konstatní a zčásti náhodné. Regulační zásah se provádí odstřelem množství zajíců nebo lišek, stanoveného regulátorem, na jehož vstupu jsou dvě regulační chyby a na výstupu dvě tříčlenné lineární kombinace těchto chyb s koeficienty nastavenými z podmínky optimálního řízení (minimální střední absolutní chyby řízení) při nulových pozorovacích chybách. Program demonstruje značnou robustnost vůči pozorovacím chybám získanou užitím gnostické regulační chyby místo klasické euklidovské.

3. Další rozvoj gnostických programů

V blízké budoucnosti budou zahájeny práce na využití gnostické teorie pro robustní korelační analýzu, pro robustní odhadování korelačních koeficientů, korelačních funkcí a korelačních matic. V další etapě by mělo dojít k ověřování možností, které gnostická teorie nabízí pro analýzu množství informace a informačních závislostí datových souborů.