OPTIMALIZACE ALGORITMŮ A ANALÝZA DAT .

Dan Pokorný

Matematické středisko biologických ústavů ČSAV
142 20 Praha 4, Vídeňská 1083

From the viewpoint of applied computational complexity, statistics is a gold mine, for it provides a rich and extensive source of unanalyzed algorithms and computational procedures. For the statistician, however, the search for efficient algorithms has not been of prime concern for several reasons: First, the design of fast algorithms is a new and developing art. Second, until recently, the cost of obtaininig data has been far greater than the cost of analysing it. Now, however, speech and image processing provide information to statistical analysis programs rapidly and cheaply, so that fast analysis is of considerable importance. Third, statisticians are properly concerned with the significance and effectiveness of the tests they perform, rather than with their cost. The result has been that the analysis of statistical algorithms remains largely ignored.

Michael Ian Shamos

Matematická statistika a analýza dat jsou jedněmi z velmi dynamicky se rozvíjejících matematických disciplin. S novými netradičními aplikacemi i teorie těchto disciplin překračuje své meze; zkoumají se data mnohorozměrná, kontaminovaná, složitě strukturovaná atd. Užitečnost nových procedur je ovšem podmíněna jejich efektivní algoritmizovatelností a konečně i existencí vhodných počítačových programů. Efektivnost procedury nerozhoduje jenom o ceně analýzy, ale často vůbec o principiální možnosti jejího nasazení. Byly to původně jiné oblasti než matematická statistika, které vyvolaly zájem o výzkum efektivních algoritmů: tak například zpracování hromadných dat stimulovalo výzkum problematiky třídění a vyhledávání, umělá inteligence přinesla problematiku prohledávání velmi rozsáhlých struktur; samozřejmě prohledávání "inteligentního". O úvodní a nepříliš systematický přehled výsledků, použitelných při konstrukci procedur analýzy dat a matematické statistiky jsme se pokusili v této přednášce.

V úvodu jsme poukázali na formalizace pojmů časové a prostorové efektivnosti, které nabízí teorie složitosti algoritmů; o praktičtěji laděné odnoži této teorie se hovoří jako o teorii konkrétních algoritmů. Viz /Aho 76/, /Gruska 74/,/Gruska 75-76/.

Jednou ze základních otázek konstrukce statistických algoritmů je třídění ve vnitřní paměti. V naší přednášce jsme na opticko-mechanickém výpočetním zařízení demonstrovali reprezentanty tří základních tříd algoritmů vnitřního třídění: třídění kvadraticky složitého /třídění bublinkové/, třídění složitosti $n \log n$ /mergesort a quicksort, který ovšem patří do této třídy pouze tak zvanou průměrnou složitostí/ a ve speciálních případech použitelného třídění s lineární složitostí /radixsort/. Úvodní informaci viz /Hořejš 80/, podrobnější /Gruska 75-76/, /Knuth 73/ a /Aho 76/.

Algoritmy rychlého vnitřního třídění jsou užitečné například pro výpočet Spearmanova koeficientu pořadové korelace; pozoruhodné je, že srovnatelně efektivně lze vypočítat i zdánlivě složitější koeficient Kendallův.

Řada dalších možných užití rychlého třídění je nasnadě, např. při výpočtu Wilcoxonova testu. Ale již při výpočtu testu mediánového lze využít značně silnějšího prostředku. Velmi překvapivým objevem posledních let bylo zjištění, že medián - i jakýkoli jiný kvantil - lze vypočítat v čase lineárním. Tedy tak složitě - až na multiplikativní konstantu - jako například výběrový průměr. Viz /Blum 72/, /Floyd 73/. Výpočet mediánového testu i v hodně rozsáhlých souborech nemusí být tedy podstatně časově náročnější než výpočet Studentova t-testu. Podle zákona o zachování obtíží je složitost přenesena na programátora: algoritmus pro rychlý výpočet výběrového mediánu je věcně dost komplikovaný. Celou řadu dalších možných algoritmů užívajících rychlého výpočtu kvantilu uvádí /Shamos 76/ v článku, z něhož jsme si jednu pasáž vypůjčili jako motto. /Např. useknutý průměr v čase n, Hodges-Lehmanův odhad v čase n log n./

Úloha opačná k určení kvantilu, tj. nalezení pořadí jednoho daného prvku je také úlohou s lineární časovou složitostí; rozdíl spočívá v trivialitě algoritmu.

Další oblastí, která může nalézt uplatnění při konstrukci statistických algoritmů, je tzv. vyhledávání. Viz /Wiedermann 81/, /Aho 76/. Jednoduchým případem vyhledávání je např. zjištění, do kterého z daných intervalů patří jistá hodnota, což lze - pomocí vhodně připravené stromové struktury - zjistit v logaritmickém čase /tj. v čase úměrném logaritmu počtu intervalů/. Tyto metody mohou najít uplatnění všude, kde se obecně spojité veličiny "kategorizují", např. při vytváření kontingenčních tabulek, při vytváření skupin v analýze rozptylu či diskriminační analýze atd.

Filosofie analýzy dat, jejímž jedním aspektem je "nabízení co nejlepšího pohledu na daná data", vede často k formulaci problémů z algoritmického hlediska velmi náročných. O řadě těchto problémů je dokonce známo, že jejich "zaručeně efektivní" algoritmizace není možná. Takové problémy existují např. v oblasti shlukové analýzy nebo v oblasti automatické formace hypotéz, jde o tzv. NP-úplné problémy.

V přednášce jsme si všimli třídy úloh, pro které je obecně charakteristické prohledávání množiny všech podmnožin jisté množiny, resp. množiny všech rozkladů. Ilustrovali jsme na příkladu mnohorozměrné lineární regrese vývoj, který v tomto případě šel od prostého počítání jednotlivých regresních vztahů přes efektivní, avšak stále ještě exhaustivní počítání všech možných lineárních regresí /Schatzoff 68/ až k "inteligentnímu" rychlému hledání nejlepších množin regresorů /Furnival 74/. Charakteristickým znakem Furnival-Wilsonova algoritmu je, že "neexhaustivním" výpočtem získáme optimální, tj. "exhaustivní" řešení.

Poslední charakteristika platí plně pro procedury, navrhované v kontextu automatické formace hypotéz; matematické základy této teorie viz /Hájek 78/. Podrobněji jsme se zabývali algoritmy pro optimální kolapsování kontingenčních tabulek.

Souvislostem mezi matematickou informatikou a matematickou statistikou je - přes citované skeptické hodnocení Shamosovo - věnována přece jenom velká i systematická pozornost; pravidelně se publikují statistické programy, pořádají mezinárodní setkání. Jiné interakce na své znepokojení patrně čekají; máme na mysli např. vztah analýzy dat a matematické statistiky s oblastí umělé inteligence. Zde jsou již činěny pokusy hledat podstatně hlubší souvislosti, než jaké jsme naznačili v naší přednášce, srv. např. /Hájek 82/.

P.S.: Množinu datových entit, se kterými jsme pracovali při demonstracích algoritmů na optico-mechanickém výpočetním zařízení, se pokusíme vydat jako přílohu k tomuto sborníku.

L I T E R A T U R A :

A. V. Aho, J. E. Hopkroft, J. D. Ullman: The design and analysis of computer programs. Addison Wesley Publishing Company 1976

M. Blum, R. W. Floyd, V. Pratt, R. Rivest, R. E. Tarjan: Time bounds for selection. JCSS 4/1972/ 448-461

COMPSTAT /zborníky konferencie/. Physica Verlag, Wien 1974, 1976, 1978, 1980

R. W. Floyd, R. Rivest: Expected time bound for selection. CACM 18/1973/

G. M. Furnival, R. W. Wilson: Regression by leaps and bounds. Technometrics 16/1974/, 449-511

J. Gruska: O zložitosti algoritmov I, II. Informačné systémy 3/1974/, 184-195, Informačné systémy 4/1975/, 335-349

J. Gruska: Zložitosť konkrétnych algoritmov /I. Základné princípy, II. Triedenie/. Učebné texty, Výskumné výpočtové stredisko, Bratislava 1975-1976.

P. Hájek, T. Havránek: Mechanizing Hypothesis Formation. Springer Verlag, Berlin - Heidelberg - New York 1978

P. Hájek, T. Havránek: GUHA 80 - an Application of AI to Data Analysis. Počítače a umelá inteligencia 1/1982/, no. 2 / v tisku/

I. M. Havel: Robotika - úvod do teorie kognitivních robotů. SNTL, Praha 1980

J. Hořejš, J. Brodský, J. Staudek: Struktura počítačů a jejich programového vybavení. SNTL-Alfa, Praha 1980

D. E. Knuth: The Art of Computer Programming, I, II, III. Addison Wesley 1976, 1977, 1978

D. Pokorný: Analýza dvourozměrných kontingenčních tabulek. In: J. Antoch /ed./: Robust 1.

M. Schatzoff, R. Tsao, S. Fienberg: Efficient Calculation of All Possible Regressions. Technometrics 10/1968/, 769-779

M. I. Shamos: Geometry and statistics: Problems at the interface. In: J. F. Traub /ed./: Algorithms and complexity, New directions and recent research. Academic Press, New York 1976, 251-280

J. Wiedermann: Vyhledávanie. In: SOFSEM 81.