JAK RYCHLE PĚSTOVAT STROMY[1]

Jan KLASCHKA a Jaromír ANTOCH

ÚIVT AV ČR a KPMS MFF UK

Abstract: This paper deals with a fast splitting algorithm for building tree-structured models. An original idea by Siciliano and Mola, who proposed a version of the algorithm for particular types of classification and regression trees, is generalized for a broad class of methods and splitting criteria. Moreover, ammount of savings of computational cost is estimated.

Резюме: В статье изучается эффективный алгорифм построения деревьев в анализе данных. Идея авторов Сицилиано и Мола для некоторых видов классификационных и регрессионных деревьев развивается таким образом, что ее обобщенная форма предложенная авторами этой статьи применима к многим методам анализа данных использующим дереввьев и к широкому классу критериев оптимальности разветвления. Обсуждаются тоже количественные аспекты економии вычислений.

Úvod aneb kde všude rostou

Pěstování stromů není jen jedním z oborů, které se vyučují na hostitelské fakultě *ROBUST* u 96. Je to také společný rys celé skupiny metod analýzy dat. Nejznámější z nich je asi *CART* popsaný v knize Breimann et al. (1984), o kterém jeden z autorů obšírně referoval na *ROBUST* u 88, blíže viz Antoch (1988). Z dalších můžeme jmenovat třeba *AID*, viz Morgan a Sonquist (1963), *FIRM*, viz Hawkins (1990), pod hlavičkou SPSS komerčně šířený *CHAID* nebo Ciampiho *RECPAM*, viz např. Ciampi (1993, 1994). Abychom čtenáře neunavovali opakovanými opisy, budeme v dalším textu pro danou metodologii užívat označení *stromové metody*.

Každá ze stromových metod při aplikaci na konkrétní problém vytváří model, jehož důležitou součástí při grafickém znázornění výsledků je orientovaný graf tvaru binárního stromu. List (koncový uzel) stromu reprezentuje nějakou podmnožinu prostoru $\mathbb{X}$ možných hodnot vektoru pozorování. Všechny listy stromu určují rozklad $\mathbb{X}$ na disjunktní podmnožiny a výsledný model je pak definován po částech odpovídajících listům.

Uveďme některé typické příklady těchto postupů.

- U klasifikačních stromů vytvořených např. metodou *CART* nebo *CHAID* je listu přiřazena určitá hodnota klasifikační funkce.
- Regresní strom vypočtený metodou *CART* nebo *AID* definuje regresní funkci, která je uvnitř každé množiny odpovídající listu konstantní.

[1] Tato práce byla podporována granty GA ČR 102/95/1311 a GA UK 188/96.

– Metoda *RECPAM* vytváří, dle typu problému, model po částech lineární regrese, po částech logistické regrese, po částech Coxův model přežívání atp.

Strom a odpovídající rozklad prostoru $\mathbb{X}$ je výsledkem rekurzivního výpočetního procesu. Začíná se kořenem[2], tj. stromem o jediném uzlu, korespondujícím s celým prostorem $\mathbb{X}$. V prvém kroku se hledá, jak $\mathbb{X}$ v jistém smyslu optimálně rozložit na dvě podmnožiny. V následujících odstavcích budeme specifikovat, jaké způsoby rozkladu přicházejí v úvahu. Zatím se spokojme s tím, že uvažovaných možností je konečně mnoho, ale obvykle velmi mnoho. Optimální rozklad $\mathbb{X}$ definuje dvojici uzlů, které se přidají ke stromu. Pro každý z těchto uzlů (resp. odpovídajících podmnožin prostoru $\mathbb{X}$) se pak opět hledá optimální rozklad na dvě části. A tak dále. Každá stromová metoda má pravidlo, podle něhož se rozhoduje, zda uzel bude listem, nebo ne (tzv. stopping rule). Toto pravidlo (je-li rozumné) mj. zaručuje, že proces přidávání uzlů je konečný.

Optimální rozklad pro daný uzel stromu se obvykle hledá způsobem principiálně jednoduchým, ale výpočetně náročným. Zpravidla se mechanicky proberou všechny možnosti, pro každou z nich se z dat vypočte určitá statistika (jaká – to závisí na tom, o kterou metodu se jedná) a vybere se ten rozklad, který tuto statistiku maximalizuje.

Počítačům (ne už tolik programátorům) může ušetřit práci, umíme-li nějakým výpočetně jednoduchým způsobem „uhodnout“, které z možných rozkladů případně mohou, a které rozhodně nemohou být optimální. Smyslem tohoto příspěvku je ukázat jeden způsob, jak v řadě stromových metod lze urychlit pěstování stromů tím, že se při optimalizaci rozkladu optimum zbytečně nehledá tam, kde nemůže být.

ROZKLAD ANEB Z ČEHO ROSTOU

Náš výklad bude z botanického hlediska poněkud zmatený. V tomto odstavci pojednáme o „kompostu“, ze kterého se rodí větve a listy (a v jistém smyslu i plody) stromů.

Zmínili jsme se již o prostoru $\mathbb{X}$ možných hodnot vektoru pozorování, zatajili jsme však, jak tento prostor vypadá. Stejně tak jsme zatím zamlčeli, jaké rozklady prostoru $\mathbb{X}$ přicházejí v úvahu.

Ve standardní situaci se u každého případu měří tytéž veličiny $X_1, X_2, \ldots, X_k$ nabývající hodnot z podmnožin $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_k$ reálné přímky, takže $\mathbb{X}$ je nějaká podmnožina $k-$rozměrného prostoru R^k. Data tedy mohou být uspořádána do klasické matice struktury *případy* $\times$ *veličiny*. Hodnotu veličiny X_i u p-tého případu, tj. hodnotu v p-tém řádku a i-tém sloupci datové matice, budeme značit $X_i(p)$.

[2] i když JK by zde raději viděl „pařez“ (pozn. JA)

Některé z veličin $X_1, X_2, \ldots, X_k$, např. v *CART*u všechny s výjimkou jedné, totiž závisle proměnné, jsou vybrány jako veličiny, jejichž hodnoty definují možné rozklady prostoru $\mathbb{X}$. Každé takové veličině X_i přísluší *množina rozkladů založených na* X_i. Co tato množina obsahuje, závisí na tom, je-li X_i veličina numerického, nebo kategoriálního typu[3].

Hodnoty veličiny numerického typu lze smysluplně uspořádat podle velikosti. Rozklady založené na veličině X_i numerického typu jsou pak dvojice množin (budeme také říkat *třídy rozkladu*) $\{X_i < c\} = \{(x_1, \ldots, x_k) \in \mathbb{X}; x_i < c\}$ a $\mathbb{X} \setminus \{X_i < c\}$, kde c probíhá množinu všech hodnot veličiny X_i v datech s výjimkou hodnoty minimální.

Rozklady založené na veličině X_i kategoriálního typu jsou tvaru $\{X_i \in B\} = \{(x_1, \ldots, x_k) \in \mathbb{X}; x_i \in B\}$ a $\mathbb{X} \setminus \{X_i \in B\}$, kde B je libovolná vlastní neprázdná podmnožina množiny všech možných hodnot X_i.

Rozklady budeme značit symbolem s. Množinu všech rozkladů založených na veličině X_i označíme $\mathbb{S}(X_i)$. Konečně $\mathbb{S}$ bude sjednocení množin $\mathbb{S}(X_i)$ přes všechny veličiny X_i určené k tomu, aby rozklady definovaly, tj. množina všech rozkladů, které přicházejí v úvahu.

Každé podmnožině prostoru $\mathbb{X}$ přirozeně odpovídá část datového souboru tvořená těmi případy p, jejichž vektor měření $\big(X_1(p), \ldots, X_k(p)\big)$ je prvkem dané podmnožiny. Bude jistě srozumitelné (aniž bychom uváděli formální definice) co znamená, že určité pozorování patří do nějakého uzlu stromu či do některé třídy rozkladu. Podobně nebudeme rozvádět, co myslíme rozkladem datového souboru nebo jeho části, je-li řeč o nějakém rozkladu prostoru $\mathbb{X}$.

Závěrem odstavce dvě okrajové poznámky. Za prvé, při výběru optimálního rozkladu pro konkrétní uzel stromu se některé z výše uvedených rozkladů mohou dostat „mimo hru". Například proto, že daná stromová metoda vyžaduje, aby do každé ze tříd rozkladu patřil alespoň určitý minimální počet pozorování (z těch případů, které patří do uzlu). Za druhé, některé stromové metody umožňují také leccos, co se uvedenému standardnímu schématu vymyká. Máme na mysli mimo jiné tyto „nadstandardní služby":

- Lze například pracovat s daty, kde u různých případů nejsou měřeny tytéž veličiny.
- Přípustné rozklady mohou být definovány zcela libovolně podle věcné podstaty problému.
- Rozklady mohou být dány nejen hodnotami jednotlivých numerických veličin, ale i hodnotami lineárních kombinací více numerických veličin atd.

Naše úvahy se těmito možnostmi stromových metod nicméně nebudou zabývat.

[3]Budeme se držet terminologie běžné v literatuře o stromových metodách, i když by se nám více líbilo mluvit o typu ordinálním a nominálním.

JAK VYBRAT NEJLEPŠÍ Z NEJLEPŠÍCH

Každá stromová metoda má kritérium (splitting rule), podle kterého se posuzuje, který rozklad je pro daný uzel stromu optimální. Popřípadě si lze vybrat mezi více takovými kritérii.

Kritérium optimality rozkladu lze zpravidla popsat následujícím způsobem. Máme nějakou statistiku φ, která závisí:

- jednak na množině případů P v tom smyslu, že se počítá z dat $\big(X_1(p), \ldots, X_k(p)\big)$, $p \in P$,
- jednak na rozkladu $s \in \mathbb{S}$.

Hodnotu této statistiky budeme značit $\varphi(P, s)$. Mějme nějaký uzel stromu a nechť P je množina všech případů z datového souboru, které do tohoto uzlu patří. Pak optimálním rozkladem pro daný uzel bude ten rozklad s, který maximalizuje $\varphi(P, s)$ přes všechna $s \in \mathbb{S}$.

Ukážeme několik příkladů. Budeme přitom vesměs předpokládat, že rozklad s prostoru $\mathbb{X}$ indukuje rozklad množiny n případů P na množiny $P_1(s)$ a $P_2(s)$ o velikosti $n_1(s)$ a $n_2(s)$.

Metoda *CART*, která řeší jak regresní, tak klasifikační úlohy, pracuje s jednou závisle proměnnou $Y = X_1$, zatímco všechny ostatní proměnné definují možné rozklady. Úloha regrese má dvě klasické varianty. V případě varianty nejmenších čtverců má statistika φ tvar

$$\varphi(P, s) = -\sum_{j=1}^{2} \sum_{p \in P_j(s)} \Big(Y(p) - \overline{Y}_j\Big)^2,$$

kde

$$\overline{Y}_j = \frac{1}{n_j(s)} \sum_{p \in P_j(s)} Y(p), \qquad j = 1, 2\,.$$

Varianta nejmenších absolutních odchylek má statistiku φ tvaru

$$\varphi(P, s) = -\sum_{j=1}^{2} \sum_{p \in P_j(s)} \big|Y(p) - \widetilde{Y}_j\big|\,,$$

kde $\widetilde{Y}_1$, resp. $\widetilde{Y}_2$, je medián hodnot $Y(p)$ v množině $P_1(s)$, resp. $P_2(s)$.

Pro klasifikační úlohu má metoda *CART* několik možných kritérií optimality rozkladu. Pro jedno z nich, kritérium založené na tzv. Giniho indexu, lze funkci φ vyjádřit následujícím způsobem. Veličina $Y = X_1$ nabývá konečně mnoha hodnot $y_1, \ldots, y_L$. Označme $n_{jl}(s)$ pro $j = 1, 2$ a $l = 1, \ldots, L$ počet pozorování z $P_j(s)$, pro která je $Y(s) = y_l$. Potom

$$\varphi(P,s) = \sum_{j=1}^{2} \sum_{l=1}^{L} \frac{n_{jl}^2(s)}{n_j(s)}.$$

Jiné stromové metody (např. *RECPAM*) pracují s modely, jejichž parametry se v každém uzlu odhadují pomocí dat, která do uzlu patří. Statistika φ pak zpravidla souvisí se ztrátovou funkcí, která je použitou metodou odhadu minimalizována.

Uvažujme např. model mnohonásobné lineární regrese pro závisle proměnnou $Y = X_1$ a nezávisle proměnné $X_2, \ldots, X_{k_0}, k_0 \leq k$. Rozklady jsou definovány pomocí veličin $X_i, k_0 < i \leq k$, ale případně i pomocí některých, či dokonce všech veličin vystupujících v modelu lineární regrese v roli nezávisle proměnných. Odhadují-li se parametry metodou nejmenších čtverců, lze ztrátu v j-té třídě rozkladu, $j = 1, 2$, vyjádřit reziduálním součtem čtverců

$$RSS\big(P_j(s)\big) = \sum_{p \in P_j(s)} \Big(Y(s) - \widehat{\beta}_1\big(P_j(s)\big) - \sum_{i=2}^{k_0} \widehat{\beta}_i\big(P_j(s)\big) X_i(p)\Big)^2,$$

kde $\widehat{\beta}_i\big(P_j(s)\big), i = 1, \ldots, k_0$ jsou odhady parametrů založené na pozorováních z $P_j(s)$. Statistika φ pak bude

$$\varphi(P,s) = -\sum_{j=1}^{2} RSS\big(P_j(s)\big)\,.$$

Obdobně lze pracovat s nějakým modelem pro vektor veličin $\boldsymbol{Y} = (X_1, \ldots, X_{k_0}), k_0 \leq k$, jehož vektor parametrů $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ se odhaduje metodou maximální věrohodnosti, např. s modelem logistické regrese. Předpokládáme-li u vektoru $\boldsymbol{Y}$ hustotu $f(\boldsymbol{y}, \boldsymbol{\theta})$ a zároveň nezávislost pozorování, je věrohodnostní funkce pro třídu rozkladu $P_j(s)$, $j = 1, 2$, dána vztahem

$$L\big(\boldsymbol{\theta}, P_j(s)\big) = \prod_{p \in P_j(s)} f\big(\boldsymbol{Y}(p), \boldsymbol{\theta}\big)$$

a její maximum je

$$ML\big(P_j(s)\big) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L\big(\boldsymbol{\theta}, P_j(s)\big).$$

Statistika φ pak bude

$$\varphi(P,s) = \prod_{j=1}^{2} ML\big(P_j(s)\big).$$

Závěrem ještě přiznáme, že schéma načrtnuté v tomto odstavci se neuplatňuje ve stromových metodách bez výjimky. Metoda *RECPAM* např. umožňuje zavést tzv. globální doprovodné proměnné. Pokud je tato možnost využita, musí být odhady některých parametrů (totiž koeficientů u globálních doprovodných proměnných) stejné ve všech listech stromu, takže výběr optimálního rozkladu pro daný uzel závisí i na datech, která do uzlu nepatří.

MUSÍ TO VŠECHNO BÝT?

Přímočarý, ale výpočetně náročný způsob maximalizace statistiky φ pro daný uzel stromu spočívá v tom, že se prostě vyzkoušejí všechny možné rozklady, tj. statistika φ se postupně vypočte pro všechna $s \in \mathbb{S}$.

Naši neapolští přátelé Roberta Siciliano a Francesco Mola přišli s nápadem, jak se dá část výpočtů ušetřit. Statistika $\varphi(P, s)$ se pro všechny rozklady $s \in \mathbb{S}(X_i)$ založené na téže veličině X_i shora odhadne číslem $\psi(P, X_i)$. Pokud známe nějaký rozklad $s^* \in \mathbb{S} \setminus \mathbb{S}(X_i)$ takový, že $\varphi(P, s^*) > \psi(P, X_i)$, nemusíme už pro rozklady $s \in \mathbb{S}(X_i)$ statistiku φ počítat, protože víme, že maximum je určitě někde jinde. Tuto myšlenku využívá tzv. algoritmus *FAST*[4], který popíšeme v následujícím odstavci.

Siciliano a Mola (1996, 1997) publikovali konstrukci horního odhadu ψ a algoritmus *FAST* využívající tento horní odhad pro několik speciálních případů stromových metod, resp. statistik φ. V této práci uvedeme obecnou formulaci podmínek, za nichž je myšlenka použitelná.

Horní odhad lze mnohdy získat relativně snadno. Nechť veličina X_i nabývá hodnot $v_1, \ldots, v_m$. Označme $s(X_i)$ rozklad prostoru $\mathbb{X}$ na m podmnožin $\{X_i = v_j\} = \{(x_1, \ldots, x_k) \in \mathbb{X}; x_i = v_j\}$. Náš horní odhad pak dostaneme jako

$$\psi(P, X_i) = \varphi\big(P, s(X_i)\big).$$

Totiž zpět. Nedostaneme, protože $s(X_i)$ je rozklad na tolik množin, kolik je hodnot veličiny X_i, zatímco statistika φ, tak jak jsme o ní dosud mluvili, je definována jen pro rozklady $s \in \mathbb{S}$, tedy vesměs rozklady na *dvě* podmnožiny. Horní odhad ψ tedy dostaneme podle uvedeného vzorce, jestliže definici funkce φ rozšíříme na obecné m-ární rozklady, $m \geq 2$. Pochopitelně se nemůže jednat o zcela libovolné rozšíření. Potřebujeme ještě jednu vlastnost, která by zaručila, že $\varphi\big(P, s(X_i)\big)$ nebude menší než $\varphi(P, s)$ pro žádný rozklad s založený na X_i.

[4]Toho, kdo by nás podezíral, že nedovedeme přeložit slovo „fast“ do češtiny, ujišťujeme, že nechceme. Siciliano a Mola toto označení zavedli jako jméno – pravda s přesvědčením, že se jedná také o výstižný přívlastek.

Řekneme, že m_1-ární rozklad s_1 prostoru $\mathbb{X}$ je zjemněním m_2-árního rozkladu s_2, jestliže každá třída rozkladu s_1 je podmnožinou některé třídy rozkladu s_2. O funkci φ řekneme, že je neklesající vůči zjemnění, pokud platí $\varphi(P, s_1) \geq \varphi(P, s_2)$, kdykoli je s_1 zjemněním s_2.

Nyní již můžeme vše upřesnit. Funkce $\psi(P, X_i)$ definovaná výše uvedeným způsobem je požadovaným horním odhadem, pokud funkce φ je dodefinována pro obecné m-ární rozklady tak, že je neklesající vůči zjemnění. Je to zřejmé z toho, že $s(X_i)$ je společným zjemněním všech binárních rozkladů z $\mathbb{S}(X_i)$.

Ukažme konstrukci horního odhadu ψ na příkladech. Použijeme k tomu funkce φ definované v předcházejícím odstavci. Budeme přitom předpokládat, že rozklad $s(X_i)$ prostoru $\mathbb{X}$ indukuje rozklad množiny n případů P na m množin $P_j\big(s(X_i)\big)$ velikosti $n_j\big(s(X_i)\big)$, $j = 1, \ldots, m$. Kde je to možné, zachováváme značení z předcházejícího odstavce.

Pro regresi metodou *CART* ve variantě nejmenších čtverců[5] je

$$\psi\big(P, X_i\big) = -\sum_{j=1}^{m} \sum_{p \in P_j(s(X_i))} \big(Y(p) - \overline{Y}_j\big)^2,$$

kde

$$\overline{Y}_j = \frac{1}{n_j\big(s(X_i)\big)} \sum_{p \in P_j(s(X_i))} Y(p), \qquad j = 1, \ldots, m.$$

Pro variantu nejmenších absolutních odchylek je

$$\psi\big(P, X_i\big) = -\sum_{j=1}^{m} \sum_{p \in P_j(s(X_i))} \big|Y(p) - \widetilde{Y}_j\big|,$$

kde $\widetilde{Y}_j$ je pro $j = 1, \ldots, m$ medián hodnot $Y(p)$ v množině $P_j\big(s(X_i)\big)$.

Pro klasifikaci metodou *CART* s využitím kritéria založeného na Giniho indexu lze funkci ψ vyjádřit jako

$$\psi\big(P, X_i\big) = \sum_{j=1}^{m} \sum_{l=1}^{L} \frac{n_{jl}^2\big(s(X_i)\big)}{n_j(s)},$$

kde $n_{jl}\big(s(X_i)\big)$ pro $j = 1, \ldots, m$ a $l = 1, \ldots, L$ je počet těch pozorování z $P_j\big(s(X_i)\big)$, pro která $Y(s) = y_l$.

[5] Pro kategoriální veličiny X_i je v programu *CART* implementován jiný způsob redukce počtu prověřovaných rozkladů. Platí, že stačí seřadit třídy rozkladu $s(X_i)$ podle velikosti průměru $\overline{Y}_j$ a pro různá c „dát k sobě“ třídy s $\overline{Y}_j \leq c$ a s $\overline{Y}_j > c$. (To se dá udělat $m - 1$ způsoby.) Binární rozklad maximalizující φ v $\mathbb{S}(X_i)$ mezi takto vytvořenými binárními rozklady jistě bude.

V situaci, kdy se v uzlech stromu odhadují parametry modelu mnohonásobné lineární regrese, máme

$$\psi(P, X_i) = -\sum_{j=1}^{m} RSS\ \Big(P_j\big(s(X_i)\big)\Big),$$

kde $RSS\big(P_j(s(X_i))\big)$ je pro $j = 1, \ldots, m$ reziduální součet čtverců dosažený při samostatném zpracování dat z $P_j\big(s(X_i)\big)$.

Konečně v případě modelu, jehož parametry se odhadují metodou maximální věrohodnosti, máme pro třídu rozkladu $P_j\big(s(X_i)\big)$, $j = 1, \ldots, m$, věrohodnostní funkci

$$L\Big(\boldsymbol{\theta}, P_j\big(s(X_i)\big)\Big) = \prod_{p \in P_j(s(X_i))} f\big(\boldsymbol{Y}(p), \boldsymbol{\theta}\big)$$

s maximem

$$ML\Big(P_j\big(s(X_i)\big)\Big) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L\Big(\boldsymbol{\theta}, P_j\big(s(X_i)\big)\Big).$$

Horní odhad ψ pak je

$$\psi(P, X_i) = \prod_{j=1}^{m} ML\Big(P_j\big(s(X_i)\big)\Big).$$

VZHŮRU NA STROMY!

Máme tedy ideu nepočítat statistiku φ pro rozklady založené na nějaké veličině, pokud horní odhad ψ ukáže, že optimální rozklad mezi nimi nemůže být, a umíme také mnohdy takový horní odhad vypočítat. Teoreticky se může tato myšlenka uplatnit u libovolné veličiny X_i, která má definovat rozklady. Na první pohled ale vidíme, že to nic nepřinese, pokud X_i nabývá jen dvou hodnot, neboť $s(X_i)$ pak obsahuje jediný rozklad a výpočet odhadu ψ je totožný s výpočtem statistiky φ pro tento rozklad. Na druhý pohled také vidíme, že idea nebude příliš užitečná ani pro numerické veličiny. Odhad ψ bude při velkém počtu hodnot veličiny X_i, a tedy velkém počtu tříd rozkladu $s(X_i)$ výpočetně relativně náročný, a přitom bude často vysoce nadhodnocený, tj. bude o příliš mnoho vyšší než maximum statistiky φ v $\mathbb{S}(X_i)$. V důsledku toho se stejně zřídka zabrání tomu, aby se statistika φ počítala pro všechny rozklady založené na X_i. Doporučujeme tedy aplikovat uvedenou myšlenku jen na kategoriální veličiny nabývající nejméně tří hodnot.

Rozdělme dále veličiny, které mají definovat rozklady, na dvě skupiny. Prvou skupinu budou tvořit veličiny kategoriální se třemi a více hodnotami

(nebo obecněji veličiny, na něž se rozhodneme aplikovat trik s horním odhadem). Veličiny v této skupině označíme $Y_1, \ldots, Y_{k_1}$. Zbylé veličiny, které budou tvořit druhou skupinu, označíme $Z_1, \ldots, Z_{k_2}$. Předpokládejme, že do nějakého uzlu t stromu patří množina pozorování P, kritérium optimality rozkladu je založeno na maximalizaci statistiky φ, a že umíme $\varphi(P, s)$ pro $s \in \mathbb{S}(Y_i)$ shora odhadnout funkcí $\psi\big(P, Y_i\big)$, $i = 1, \ldots, k_1$.

Algoritmus *FAST*, resp. jeho pražská varianta, potom bude pracovat takto.

(1) Jestliže $k_2 > 0$, vypočte se

$$\varphi^* = \max_{i=1,\ldots,k_2} \max_{s \in \mathbb{S}(Z_i)} \varphi\big(P, s\big)$$

a za s^* se vezme (některý) rozklad, pro který $\varphi\big(P, s^*\big) = \varphi^*$.
Pokud $k_2 = 0$, položí se $\varphi^* = -\infty$ a s^* se nedefinuje.

(2) Pro $i = 1, \ldots, k_1$ se vypočte $\psi\big(P, s(Y_i)\big)$. Veličiny $Y_1, \ldots, Y_{k_1}$ se uspořádají do posloupnosti $Y_{(1)}, \ldots, Y_{(k_1)}$ tak, aby pro $i > j$ platilo $\psi\big(P, s(Y_{(i)})\big) \geq \psi\big(P, s(Y_{(j)})\big)$.

(3) Položí se $i = 1$.

(4) Pokud je $i > k_1$, nebo pokud je $i \leq k_1$ a platí $\psi\big(P, Y_{(i)}\big) \leq \varphi^*$, přijme se s^* za optimální rozklad pro uzel t a výpočet končí.

(5) Vypočte se

$$\varphi_i^* = \max_{s \in \mathbb{S}(Y_{(i)})} \varphi\big(P, s\big)$$

a za s_i^* se vezme (některý) rozklad, pro který $\varphi\big(P, s_i^*\big) = \varphi_i^*$.
Pokud $\varphi_i^* > \varphi^*$, položí se $s^* = s_i^*$ a $\varphi^* = \varphi_i^*$.

(6) Index i se zvýší o 1 a algoritmus se vrací k bodu 4.

Princip algoritmu *FAST*, jak vidíme, je jednoduchý. Rozklady se probírají jeden za druhým, dokud se neukáže, že mezi zbylými rozklady s ohledem na horní odhady ψ už nic k nalezení není. Děje se tak samozřejmě v rozumném pořadí, které dává naději, že se výpočet zastaví relativně brzy.

POČÍTÁ, CO MU TO ...

V tomto odstavci se budeme věnovat spekulacím o tom, kolik výpočtů může aplikace algoritmu *FAST* ušetřit.

V prvním přiblížení můžeme úsporu měřit počtem rozkladů, pro které se nemusí počítat statistika φ. Bude to zřejmě záviset na mnoha okolnostech. Mimo jiné na tom, jak „strmý“ či „plochý“ průběh bude funkce φ mít, tj. zda vysokých hodnot statistiky φ dosahují jen rozklady založené na několika málo veličinách, nebo zda je mnoho rozkladů s hodnotami φ blízkými maximu. Za samostatnou studii by stála také otázka, jak mnoho (a v závislosti

na čem) odhad ψ nadhodnocuje maximum φ pro rozklady založené na určité veličině; úspora zjevně bude tím větší, čím je horní odhad ψ „těsnější“. Zde tyto problémy analyzovat nebudeme. Siciliano a Mola (1996) aplikovali algoritmus *FAST* v úloze shodné s variantou regrese metodou nejmenších čtverců v metodě *CART*. Při zpracování reálných dat dosáhli úspory cca 75 procent, tj. bylo nutno „propočítat“ jen asi každý čtvrtý rozklad.

Je tu ale první **ale**: *Výpočet horní hranice ψ také něco stojí.* Předpokládejme pro jednoduchost, že zhruba tolik, co výpočet statistiky φ pro jeden rozklad. Čtenář může posoudit, nakolik realistický je tento předpoklad v situaci dříve uvedených příkladů.

V dalším se omezíme na zkoumání úspory výpočtů týkajících se rozkladů založených na jedné pevné kategoriální veličině nabývající m hodnot. Počet různých binárních rozkladů založených na takové veličině je $r(X) = 2^{m-1} - 1$. V každém uzlu stromu, který není bez výpočtů prohlášen za list, nastane jedna ze dvou variant.

- Buďto se bude počítat jen horní odhad ψ, což je dle předpokladu ekvivalentní výpočtu φ pro jeden rozklad.
- Nebo se kromě ψ bude muset počítat také φ pro $r(X)$ rozkladů.

Druhá varianta je tedy výpočetně $\big(r(X)+1\big)$-krát dražší. S velkou licencí budeme na druhou variantu pohlížet jako na náhodný jev s pravděpodobností $\pi(X)$. Kdykoli to budeme v dalším textu potřebovat, budeme navíc předpokládat, že stejná pravděpodobnost $\pi(X)$ „funguje“ současně ve všech uzlech stromu.

Střední výpočetní náklady spojené s rozklady založenými na X pak lze vyjádřit jako

$$cost_{FAST} = 1 - \pi(X) + \pi(X)\big(r(X)+1\big) = 1 + \pi(X)r(X).$$

Pokud se algoritmus *FAST* nepoužije, bude se s pravděpodobností jedna počítat $r(X)$ hodnot statistiky φ, tj. náklady budou

$$cost_{BEZ} = r(X).$$

Relativní úspora pak je

$$RS = \frac{cost_{BEZ} - cost_{FAST}}{cost_{BEZ}} = 1 - \pi(X) - \frac{1}{r(X)}\cdot$$

Všimněme si, že pro $m = 2$ nic nevyděláme. Pro $m = 3$ algoritmus spíše pomáhá než škodí pro $\pi(X) < 2/3$. Pro větší m se relativní úspora neliší dramaticky od $1-\pi(X)$, tedy hodnoty, k níž dojdeme, ignorujeme-li náklady na výpočet horního odhadu ψ. První **ale** tedy není zvlášť závažné[6].

[6]Počet hodnot veličiny X se ovšem do úspor promítne ještě jinak, než je ze vzorců na první pohled patrné. Bude totiž mít pravděpodobně vliv i na to, jak nadhodnocený je horní odhad ψ. To zřejmě ovlivní i hodnotu $\pi(X)$.

Druhé **ale** bude daleko horší. Dosavadní analýza úspor vycházela z toho, že náklady na výpočet horní hranice ψ a na výpočet statistiky φ pro jednotlivé rozklady jsou vesměs aditivní. To bude platit tehdy, když každá z těchto hodnot se počítá zvlášť z dat. Najdou se jistě situace, kdy se nic lepšího dělat nedá. Pokud víme, je tomu tak např. tehdy, když se v uzlech odhadují parametry modelu logistické regrese, přičemž model obsahuje spojité nezávisle proměnné veličiny. Často ale lze výpočty optimalizovat tak, že se z dat vypočte malý počet údajů, z nichž se pak už dá odvodit všechno ostatní: jak ψ, tak všechny hodnoty φ. Těmto údajům budeme říkat *pomocné statistiky*.

Vrátíme-li se k příkladům z předminulého odstavce, vidíme, že tento způsob výpočtů využívající pomocných statistik se hodí pro následující tři případy.

- V regresní úloze řešené metodou *CART* ve variantě nejmenších čtverců stačí z dat spočítat pro jednotlivé třídy $P_j\big(s(X)\big), j = 1, \ldots, m$ rozkladu $s(X)$ statistiky

$$\sum_{p \in P_j\,(s(X))} Y(p) \text{ a } \sum_{p \in P_j\,(s(X))} Y^2(p)$$

  a počty pozorování.
- V klasifikaci metodou *CART* s využitím kritéria založeného na Giniho indexu je obdobně vše další funkcí frekvencí $n_{jl}, j = 1, \ldots, m,\ l = 1, \ldots, L$.
- V situaci, kdy se v uzlech odhadují parametry modelu mnohonásobné lineární regrese, stačí ke všem dalším výpočtům pomocné statistiky pro třídy $P_j\big(s(X)\big), j = 1, \ldots, m$ rozkladu $s(X)$, jak je vidět z následující úvahy. Odhad $\widehat{\boldsymbol{\beta}}$ metodou nejmenších čtverců v regresní úloze $\boldsymbol{y} = \boldsymbol{U}\beta + \boldsymbol{e}$ je $\widehat{\boldsymbol{\beta}} = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{y}$, tj. je určen maticí $\boldsymbol{U}'\boldsymbol{U}$ a vektorem $\boldsymbol{U}'\boldsymbol{y}$. Týmiž statistikami je určen i reziduální součet čtverců. Jsou-li k dispozici tyto statistiky vypočtené zvlášť pro několik nepřekrývajících se datových souborů, lze analogické statistiky pro soubor vytvořený spojením těchto souborů získat jako součet statistik pro jednotlivé dílčí soubory.

Případ regrese metodou *CART* ve variantě nejmenších absolutních odchylek je příkladem situace „někde mezi“. Znalost mediánu a součtu absolutních odchylek od mediánu ve třídách $P_j\big(s(X)\big), j = 1, \ldots, m$, rozkladu $s(X)$ by se zřejmě dala využít k poněkud efektivnějšímu výpočtu obdobných charakteristik pro třídy binárních rozkladů založených na X, než je výpočet „od nuly“.

Po těchto přípravách můžeme konečně přikročit k jádru našeho druhého **ale**. Uvažujme situaci, kde je efektivní využívat výše naznačeným způsobem pomocné statistiky. Stromové metody se typicky aplikují na velké datové soubory. Jsou-li výpočty dobře organizovány – a jen tehdy má zřejmě smysl chlubit se dalším zvýšením efektivity – padne naprostá většina výpočetních

nákladů na výpočet pomocných statistik, a to stejných statistik, ať se algoritmus *FAST* aplikuje, nebo ne. Podtrženo a sečteno, *v uvedených situacích nepřinese algoritmus FAST prakticky nic.*

Naštěstí máme v záloze ještě třetí **ale**, **kontra-ale** proti **ale** druhému. Naše úvaha byla přeci jen povrchní. Stromové metody se sice typicky aplikují na velké datové soubory, ale do některých uzlů stromu patří málo pozorování – a takových uzlů je mnohdy většina. Abychom to objasnili, musíme něco více říci o stromových metodách.

Strom, který je konečným výsledkem aplikace stromové metody na reálná data, má často jen několik málo uzlů. Některé stromové metody, zvláště starší (např. *AID*), tento výsledný strom vytvářejí relativně jednoduše. Mají dosti přísná pravidla určující, kdy je přidání dalšího uzlu ještě dostatečným přínosem. Ke stromu se přidávají uzly, dokud další růst není zakázán; pak je výsledný strom hotov. V současnosti se za lepší považuje jiný, složitější a výpočetně náročnější přístup. V prvé fázi se záměrně uplatňují velmi mírná pravidla pro zastavení růstu. Typicky se za list prohlásí uzel, do kterého patří méně než 5 či 10 pozorování. Tímto způsobem se vypěstuje tzv. *velký strom*. V další fázi se velký strom *prořezává*, tj. některé (zpravidla skoro všechny) uzly se odstraňují. Teprve prořezáváním vznikne strom přiměřené velikosti (right-sized tree, honest tree), který je konečným výsledkem aplikace metody. Při prořezávání se běžně pěstují další pomocné velké stromy. Protože se nyní zabýváme výpočetními aspekty stromových metod, není pro nás tak důležité, jak vypadá výsledný strom po prořezávání, ale co se děje při pěstování velkého stromu. A o velkých stromech skutečně skoro vždy platí, že do většiny jejich uzlů patří málo pozorování.

Pokusme se tedy o hrubou kvantifikaci úspor při použití algoritmu *FAST* v situaci, kdy se efektivní výpočty opírají o pomocné statistiky. Budeme opět uvažovat kategoriální veličinu X nabývající m hodnot a budeme odhadovat výpočetní náklady spojené s rozklady založenými na X *v celém procesu pěstování velkého stromu*. Pro jednoduchost budeme studovat jen případ, kdy náklady výpočtu pomocných statistik jsou přímo úměrné počtu pozorování v uzlu. Zavedeme parametr b, který udává, kolikrát náročnější je vypočítat z pomocných statistik jednu hodnotu φ nebo ψ než zpracovat jedno pozorování při výpočtu pomocných statistik.

Náklady na výpočty v uzlu t, do kterého patří N_t pozorování, budou bez použití algoritmu *FAST*

$$cost_{BEZ}(t) = N_t + b\, r(X),$$

zatímco s jeho použitím dostaneme

$$cost_{FAST}(t) = N_t + b\left(1 + \pi(X) r(X)\right).$$

Pro celý strom pak platí

$$cost_{BEZ} = \nu + b\, r(X)\tau$$

a

$$cost_{FAST} = \nu + b\,\big(1 + \pi(X)r(X)\big)\tau\,,$$

kde τ je počet uzlů, kterých se výpočty týkají, tj. zpravidla všech uzlů kromě listů, a ν je celkový počet pozorování ve všech těchto uzlech. Relativní úspora tentokrát vychází po jednoduchých úpravách jako

$$RS = \frac{cost_{BEZ} - cost_{FAST}}{cost_{BEZ}} = \frac{1 - \pi(X) - \dfrac{1}{r(X)}}{1 + \dfrac{\nu/\tau}{b\, r(X)}}\,.$$

Je zřejmé, že poroste-li počet pozorování v souboru nade všechny meze, poroste podíl ν/τ do nekonečna. Jinými slovy, při zpracování nesmírně velkých souborů aplikace algoritmu *FAST* skutečně skoro nic neušetří (viz **ale** č. 2). Prakticky zajímavé ale je, *jak rychle* podíl ν/τ do nekonečna roste. Na to nemáme univerzální odpověď. Pro strom na obr. 1 platí $\nu/\tau \approx n \log_2 \frac{N}{n}$, což by mohlo být „k životu". Naproti tomu pro „zvrhlý" strom na obr. 2 $\nu/\tau \approx \frac{N}{2}$ – škoda slov.

Algoritmus *FAST* tedy může být k něčemu i k ničemu. Příklady naznačují, že podstatné bude, jak symetricky či asymetricky budou optimální rozklady v uzlech stromu rozdělovat pozorování. Dokonalá symetrie vede k uspokojivým výsledkům, krajní asymetrie ke katastrofě.

Pokusme se přiblížit jakémusi „spravedlivému středu" mezi krajnostmi znázorněnými na obr. 1 a 2. Představme si, že počty pozorování v uzlech splňují následující předpoklady.

- Uzel je listem, právě když do něj patří nejvýše n pozorování.
- Nechť uzel t obsahuje $N_t > n$ pozorování, která tvoří množinu P, a optimální rozklad P vytvoří podmnožiny P_1 a P_2. Potom počet pozorování v P_1 je realizací náhodné veličiny s rovnoměrným rozdělením na $\{1, 2, \ldots, N_t - 1\}$.

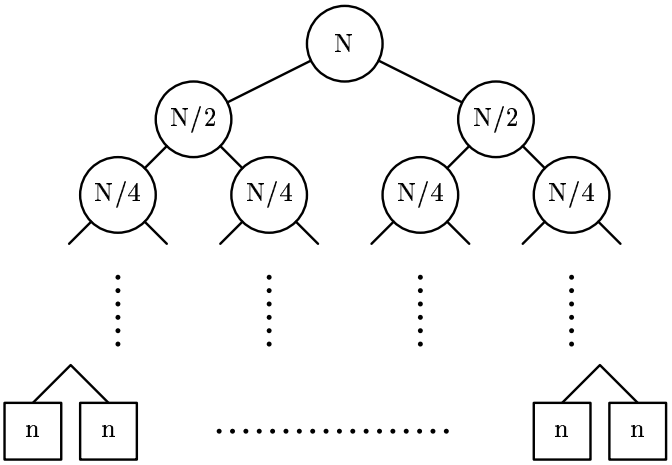
Lze snadno odvodit, že potom

$$\nu/\tau \approx (n+1) \sum_{i=n+1}^{N} \frac{1}{i} \approx (n+1) \ln \frac{N}{n}\,.$$

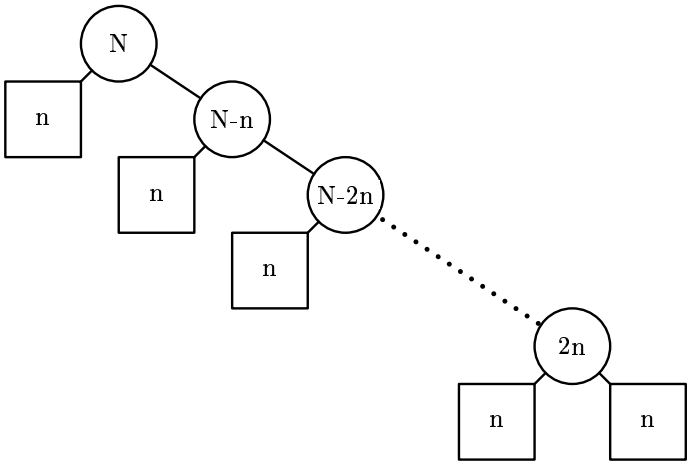
Závislost na N je podobně jako v příkladu z obr. 1 „jen" logaritmická. Je to pouze dílčí výsledek, přece jen však v nás budí umírněný optimismus.

Můžeme shrnout:

- Ušetříme o něco méně, než se zdálo, když jsme ignorovali možnost pracovat s pomocnými statistikami.
- Úspora je, což nepřekvapí, tím větší, čím více hodnot veličina X nabývá (ale pozor na současný vliv na $\pi(X)$!), čím náročnější jsou výpočty z pomocných statistik ve srovnání s výpočty z dat, a zpravidla také tím, čím méně pozorování je v jednotlivých listech.



Obr. 1. V uzlech vyznačeny počty pozorování.



Obr. 2. V uzlech vyznačeny počty pozorování.

– Úspora se zmenšuje s rostoucím celkovým počtem pozorování N. Zatímco ale tento pokles bude často velmi pomalý, závislost na počtu pozorování v listech bude daleko výraznější.

Závěrem považujeme za nutné připustit, že o skutečných úsporách rozhodne jeden důležitý strom, který je, jak praví klasik, na rozdíl od naší teorie zelený.

POROSTOU NÁM STROMY RYCHLEJI?

Autorům v nejbližší době sotva – nechystají se momentálně žádnou stromovou metodu programovat. Tomu, kdo se chystá, by nicméně rychleji růst mohly. A důvodů pro programátorské hrátky se stromy se zatím najde dost.

– Vývoj metod není uzavřený (k radosti těch, kdo o nich chtějí bádat, k zoufalství toho, kdo je potřebuje použít; vyzkoušeno obé).
– Co je popsáno v literatuře, často není v komerčních verzích software (viz *CART* a uživatelem definovaná kritéria optimality rozkladu). Někdy dokonce komerční verze vůbec chybí (viz *RECPAM*).
– To, co se tváří jako standardní program, může být soubor chyb (viz stromy v S+; vděčně vzpomínáme a „děkujeme“ za ztrátu času).
– Atd. atd.

Takže hodně zdaru.

Nemůžeme se ale ubránit jistým pochybnostem. Nebylo by bývalo přeci jen lepší pěstovat stromy na zahradě a čekat?

Zde končí text, na kterém byli autoři schopni se shodnout. Jejich představy o tom, na co by snad bývalo bylo lepší čekat a proč, se natolik rozešly, že nastal rozklad autorského týmu. Nezbylo proto, než závěr definovat po částech.

JK
V době, kdy jsme se začali o problematiku tohoto článku zajímat, byl procesor Intel Pentium®100 poslední výkřik.

JA
Ostatně kolegové v Lednici to umí, podobně jako víno, docela obstojně. Já budu asi raději čekat na to víno.

LITERATURA

Ahn H., *Log-normal regression modelling through recursive partitioning*, Computational Statistics and Data Analysis **21** (1996), 381 – 398.

Antoch J., *Klasifikace a regresní stromy*, ROBUST'88, Praha, 1988, pp. 0 – 6.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and Regression Trees*, Wadsworth, Belmont CA, 1984.

Chambers J.M., Hastie T.J., *Statistical Models in S-Plus*, Wadsworth, Belmont CA, 1992.

Chou P.A., *Optimal partitioning for classification and regression trees*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13(4)** (1991), 340 – 353.

Ciampi A., *Constructing prediction trees from data, the RECPAM approach*, Computational Aspects of Model Choice (Antoch J., ed.), Physica – Verlag, Heidelberg, 1993, pp. 105 – 156.

Ciampi A., *Classification and discrimination, the RECPAM approach*, COMPSTAT'94 (Dutter R., Grossmann W., eds.), Physica – Verlag, Heidelberg, 1994, pp. 129 – 147.

Ciampi A., Hendricks L., Lou Z., *Discriminant analysis for mixed variables: Integrating trees and regression models*, Multivariate Analysis: Future Directions (Rao C.R., Cuadras C.M., eds.), North Holland, Amsterdam, 1993, pp. 3 – 22.

Gelfand S.B., Ravishankar C.S., Delp E.J., *An iterative growing and pruning algorithm for classification tree design*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13(2)** (1991), 163 – 174.

Hawkins D.M., *FIRM (Formal Inference-based Recursive Modeling)*, Technical Report Number 546, University of Minnesota, School of Statistics, 1990.

Kass G.V., *An exploratory technique for investigating large quantities of categorical data*, Applied Statistics **24(2)** (1980), 119 – 127.

Keprta S., *Non-binary classification trees*, Statistics and Computing (1996), 231 – 243.

Mingers J., *Empirical comparison of selection measures for decision tree induction*, Machine Learning **3** (1989), 319 – 342.

Mola F., Siciliano R., *A two-stage predictive splitting algorithm in binary segmentation*, Computational Statistics (Dodge Y., Whittaker J., eds.), Physica – Verlag, Heidelberg, 1992, pp. 179 – 184.

Mola F., Siciliano R., *Alternative strategies and CATANOVA testing in two-stage binary segmentation*, New Approaches in Classification and Data Analysis (Diday E., ed.), Springer, Heidelberg, 1994, pp. 316 – 323.

Mola F., Siciliano R., *A fast splitting procedure for classification trees*, Statistics and Computing (1997) (to appear).

Morgan J.N., Sonquist J.A., *Problems in the analysis of a survey data, and a proposal*, JASA **58** (1963), 415 – 434.

Quinlan J.R., *Induction of decision trees*, Machine Learning **1(1)** (1986), 81 – 106.

Quinlan J.R., Rivest R.L., *Inferring decision trees using the minimum description length principle*, Information and Computing **80** (1989), 227 – 248.

Safavian S.R., Landgrebe D., *A survey of decision tree classifiers*, IEEE Transactions on Systems, Man, and Cybernetics, **21(3)** (1991), 660 – 674.

Siciliano R., Mola F., *A two-stage predictive splitting algorithm in binary segmentation*, COMPSTAT 94 (Dutter R., Grossmann W., eds.), Physica – Verlag, Heidelberg, 1994, pp. 172 – 177.

Siciliano R., Mola F., *A fast regression trees procedure*, Statistical Modeling, Proceedings of the 11th International Workshop on Statistical Modeling (Forcina A. et al., eds.), Graphos, Perugia, 1996, pp. 332 – 340.

Venables W.N., Ripley B.D., *Modern Applied Statistics with S-Plus*, Springer, Heidelberg, 1994.

Wallace C.S., Patrick J., *Coding decision trees*, Machine Learning **11** (1993), 7 – 22.