Změkčování rozhodovacích stromů maximalizací plochy pod částí ROC křivky

Dvořák, Jakub
2008

Změkčování rozhodovacích stromů maximalizací plochy pod částí ROC křivky

doktorand:
MGR. JAKUB DVOŘÁK
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
dvorak@cs.cas.cz

školitel:
RNDR. PETR SAVICKÝ, CSC.
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
savicky@cs.cas.cz

obor studia:
Teoretická informatika

Tento výzkum byl podporován institucionálním výzkumným záměrem AV0Z10300504 a také projektem T100300517 programu ,,Informační společnost" AV ČR.

Abstrakt

V návaznosti na plochu pod ROC křivkou jakožto obvyklou míru kvality klasifikátoru zavádíme plochu pod počáteční částí ROC křivky, která je mírou kvality klasifikátoru zaměřeného na dosažení nízké chybovosti na negativních (background) případech. Tato míra je použita jako cílová funkce při změkčování rozhodovacích stromů pomocí optimalizace. Pro optimalizaci je použit algoritmus Nelder-Mead. Experimenty na datech „Magic Telescope" ukazují účinnost této metody.

1. Úvod

Změkčování hran v rozhodovacích stromech umožňuje zlepšení klasifikátoru při zachování většiny dobrých vlastností rozhodovacích stromů. Změkčené stromy oproti klasickým mohou dosahovat lepšího poměru správné / chybné klasifikace a dalším přínosem je spojitost výstupu klasifikátoru. Zachována zůstává snadná interpretovatelnost modelu a přímočará převoditelnost na systém pravidel (v případě změkčeného stromu půjde o fuzzy-pravidla). Nevýhodou je zvětšení paměťové náročnosti modelu a hlavně časové složitosti jak učení, tak klasifikace.

Zde se budeme zabývat změkčováním jakožto postprocessingem stromů získaných standardní metodou CART [2]. Základní tvar změkčení je stejný, jako je v metodě C4.5 [6], ale liší se způsob určení (učení) parametrů, tj. hranic intervalů změkčení. Zatímco C4.5 určuje parametry změkčení pomocí směrodatné odchylky klasifikační chyby nezměkčeného stromu bez ohledu na to, jaký efekt má změkčení na chování klasifikátoru, my budeme hledat změkčení pomocí optimalizace výsledků změkčeného stromu.

Při změkčování pomocí optimalizace je pro kvalitu výsledného klasifikátoru i pro rychlost optimalizace zásadní volba cílové funkce. Použití relativního počtu chybných klasifikací se ukázalo jako nevhodné, protože to je funkce po částech konstantní a má velké množství lokálních minim. Varianty založené na sumaci transformované diference spojitého výstupu klasifikátoru a očekávané klasifikace pomohou získat spojitou funkci, ale stále trpí problémem lokálních minim a pro jejich optimalizaci byla používána metoda založená na simulovaném žíhání, jak bylo popsáno v [3], tento algoritmus je však časově velmi náročný.

V tomto příspěvku ukážeme využití plochy pod počáteční částí ROC křivky jakožto cílové funkce pro optimalizaci změkčení rozhodovacího stromu. Ukazuje se, že pro takovouto optimalizaci je možné použít simplexový algoritmus (Nelder-Mead) [5], což vede k podstatně rychlejšímu učení, než předchozí přístup se simulovaným žíháním.

2. ROC křivka a plocha pod křivkou

ROC křivka (Receiver Operating Characteristic curve) je standardním nástrojem pro analýzu chování klasifikátoru. V této sekci uvádíme především informace podstatné pro další vysvětlení změkčování rozhodovacích stromů. Čerpáme zejména z [4] a další literatury.

Pro klasifikátor, který rozděluje data do dvou tříd (nazývejme je pozitivní a negativní, někdy též signal resp. background), ROC křivka ukazuje vztah relativního počtu správně klasifikovaných pozitivních vzorů a relativního počtu chybně klasifikovaných negativních vzorů (signal acceptance vs. background acceptance) při různě nastavené „citlivosti".

Pokud výstupem klasifikátoru je pro každý datový vzor $\mathbf{x}$ reálné číslo „response" $\mathrm{R}(\mathbf{x})$, přičemž jeho vyšší hodnota reprezentuje vyšší pravděpodobnost, že předložený případ je pozitivní, potom různé nastavení citlivosti odpovídá různým volbám hodnoty prahu, kterým oddělujeme případy, jež podle response považujeme za pozitivní od případů, které zařadíme k negativním.

Plocha pod ROC křivkou (Area Under Curve, AUC) je skalárním vyjádřením kvality klasifikátoru. AUC klasifikátoru, který zařadí všechny vzory správně, je rovna jedné. Čím je hodnota nižší, tím je klasifikátor horší. AUC pro náhodný klasifikátor je $1/2$. Hodnoty v intervalu $\langle 0, 1/2)$ by charakterizovaly klasifikátor horší než náhodný.

Máme-li množinu, jež obsahuje P pozitivních vzorů $\mathbf{x}_1^+, \ldots, \mathbf{x}_P^+$ a Q negativních vzorů $\mathbf{x}_1^-, \ldots, \mathbf{x}_Q^-$ a definujeme-li funkci

$$g(u,v) = \begin{cases} 1 & \text{když } u > v \\ 1/2 & \text{když } u = v \\ 0 & \text{když } u < v \end{cases}$$

potom z této množiny vypočteme

$$\mathrm{AUC} = \frac{1}{PQ} \sum_{i=1}^{P} \sum_{j=1}^{Q} g\left(\mathrm{R}(\mathbf{x}_i^+), \mathrm{R}(\mathbf{x}_j^-)\right)$$

3. Metoda změkčování

Mějme nezměkčený rozhodovací strom, který pro vstupní vzor $\mathbf{x} = (x_1, \ldots, x_m)$ testuje ve vnitřních uzlech $v_j, j = 1, \ldots, s$ podmínky tvaru

$$x_{k_j} \leq c_j \tag{1}$$

V listech jsou uloženy hodnoty response z intervalu $\langle 0, 1\rangle$. Klasifikace tímto stromem probíhá tak, že pro předložený vzor se počínaje kořenem stromu testuje nerovnost (1), je-li splněna, pokračuje se v levém podstromu, jinak v pravém podstromu, dokud není dosaženo listu, který určí výslednou response.

Odpovídající změkčený strom bude mít stejnou strukturu, hodnoty response v listech zůstanou stejné, ale každý vnitřní uzel bude kromě hodnot k_j, c_j z podmínky (1) určovat reálné parametry změkčení $a_j, b_j \geq 0$. Potom definujeme změkčující funci f_j jež lineárně interpoluje body uvedené v tabulce:

t	$-\infty$	$-a_j$	0	b_j	∞
$f_j(t)$	1	1	1/2	0	0

Response změkčeného stromu je definována rekurzívně: v listu stromu je pro libovolný vstupní vzor response daná hodnotou uloženou v tomto listu. Jinak pro strom s kořenem v_j a vzor $\mathbf{x}$ je výsledkem průměr response levého a pravého podstromu vážený hodnotami $r_{j,\mathbf{x}}$ a $(1 - r_{j,\mathbf{x}})$, kde $r_{j,\mathbf{x}} = f_j(x_{k_j} - c_j)$.

Úlohou změkčování je pak určení parametrů $a_j, b_j, j = 1, \ldots, s$, k čemuž používáme optimalizaci funkce založené na tom, jak změkčený strom s danými parametry klasifikuje vzory z trénovací množiny.

V mnohých skutečných klasifikačních úlohách (včetně klasifikace dat „Magic Telescope" použitých v našich experimentech), je podstatné dosažení nízké úrovně background acceptance. Protože background acceptance tvoří horizontální osu ROC křivky, charakterizuje chování klasifikátoru při nízkých hodnotách background acceptance počáteční část ROC křivky. Naše metoda proto používá jako cílovou funkci pro optimalizaci plochu pod nejmenší částí ROC křivky, jež pokrývá celou oblast, kde background acceptance není větší, než zvolená hodnota $0 \leq \Theta \leq 1$. Tuto částečnou AUC označujme AUC_Θ.

Předpokládejme dále bez újmy na obecnosti, že vzory v množině, z níž počítáme AUC_Θ, jsou očíslovány tak, aby

$$\mathrm{R}(\mathbf{x}_1^+) \geq \mathrm{R}(\mathbf{x}_2^+) \geq \cdots \geq \mathrm{R}(\mathbf{x}_P^+)$$
$$\mathrm{R}(\mathbf{x}_1^-) \geq \mathrm{R}(\mathbf{x}_2^-) \geq \cdots \geq \mathrm{R}(\mathbf{x}_Q^-)$$

Označme ϑ nejvyšší hodnotu prahu, při níž je hodnota background acceptance alespoň Θ:

$$\vartheta = \mathrm{R}(\mathbf{x}^-_{\lceil \Theta Q \rceil})$$

Dále počty pozitivních a negativních případů, jejichž response je alespoň ϑ označme:

$$P_\vartheta = \max\left\{i; \quad \mathrm{R}(\mathbf{x}_i^+) \geq \vartheta\right\}$$

$$Q_\vartheta = \max\left\{j; \quad \mathrm{R}(\mathbf{x}_j^-) \geq \vartheta\right\}$$

Potom

$$\mathrm{AUC}_\Theta = \frac{1}{PQ} \sum_{i=1}^{P_\vartheta} \sum_{j=1}^{Q_\vartheta} g\left(\mathrm{R}(\mathbf{x}_i^+), \mathrm{R}(\mathbf{x}_j^-)\right)$$

Tato hodnota je vypočtena lehce modifikovaným algoritmem pro výpočet standardní AUC uvedeným v [4].

Pro optimalizaci cílové funkce je použit simplexový algoritmus pro minimalizaci (Nelder-Mead) [5]. Minimalizuje se $-\mathrm{AUC}_\Theta$ vypočtená z trénovacích dat. Algoritmus vyžaduje, aby ve vstupním prostoru měly všechny dimenze stejnou škálu, tedy aby jednotkový

krok v libovolném směru měl vždy přibližně stejný význam. Použitá škála byla definována následovně: Nejprve celý prostor ve všech směrech omezíme nejzazšími trénovacími vzory, tak získáme základní hyperkvádr. Když v uzlu v_j podmínka (1) rozděluje hyperkvádr vyšší úrovně, který je v proměnné x_{k_j} omezen hodnotami $z_{j,1}, z_{j,2}$, kde $z_{j,1} < c_j < z_{j,2}$, potom za jednotkový krok v parametru a_j resp. b_j považujeme $c_j - z_{j,1}$ resp. $z_{j,2} - c_j$. Zároveň jako iniciální hodnoty parametrů pro změkčování se použijí:

$$a_j^0 = \frac{1}{4}(c_j - z_{j,1}); \quad b_j^0 = \frac{1}{4}(z_{j,2} - c_j)$$

4. Výsledky experimentů

Pro experimenty byla použita data „Magic Telescope"[1], která jsou zkoumána také v [1] a [3]. Trénovací množina obsahovala 12680 vzorů, byla rozdělena na dvě části v poměru velikostí 2:1, první část byla použita pro růst stromu a druhá část jako validační množina pro prořezávání. Strom byl vytvořen metodou CART, velikost stromu je možno řídit nastavením parametrů prořezávání (viz [2]).

Pro změkčení byla použita výše popsaná metoda s parametrem $\Theta = 1/10$, jakožto data pro výpočet částečné AUC byla použita celá trénovací množina. Pro hodnocení získaného klasifikátoru byla použita testovací množina o velikosti 6340 vzorů.

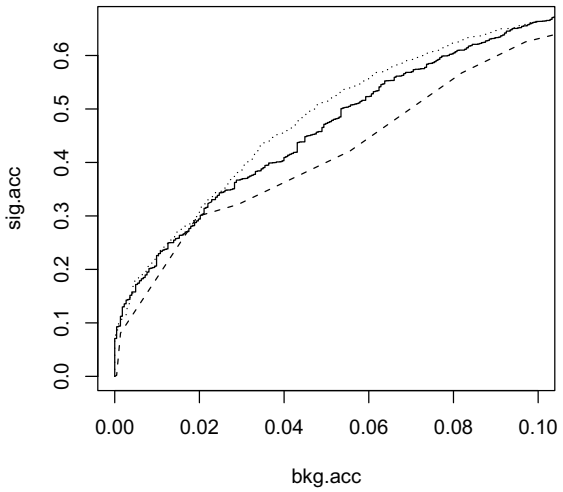
Obrázky 1 a 2 ukazují získané části ROC křivek pro vybrané stromy. Na obrázcích je čárkovaně vyznačena ROC křivka nezměkčeného stromu na testovacích datech; tečkovaná je ROC křivka změkčeného stromu na trénovacích datech, tzn. jedná se o křivku, která figurovala v cílové funkci; plnou čarou je ROC křivka změkčeného stromu na testovacích datech.

Z obrázků je patrné, že změkčený strom je v oblasti nízké úrovné background acceptance lepší klasifikátor, než nezměkčený strom.Takové chování se ukázalo jako typické i na dalších stromech.

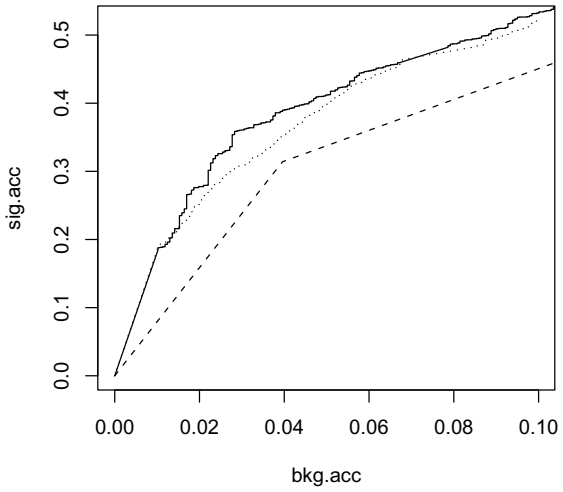
5. Závěr

Plocha pod částí ROC křivky se ukazuje jako vhodná cílová funkce pro změkčování rozhodovacích stromů pomocí optimalizace. Tuto cílovou funkci lze optimalizovat metodou Nelder-Mead, což proti doposud zkoumaným cílovým funkcím optimalizovaným pomocí simulovaného žíhání vede k významnému snížení časové náročnosti změkčování. Dalším přínosem je možnost preferovat nízkou background acceptance klasifikátoru.

V dalším výzkumu se zaměříme na ladění parametrů optimalizačního algoritmu a budeme ještě zkoumat modifikace cílové funkce. Pozornost bude také věnována skutečnosti, že v provedených experimentech byla na menších stromech zkoumaná část ROC křivky vypočtené z testovacích dat lepší, než ROC z trénovacích dat. Tento aspekt je viditelný na obrázku 2 a byl pozorován i na dalších stromech.



Obrázek 1: Části ROC křivek pro strom se 45 vnitřními uzly



Obrázek 2: Části ROC křivek pro strom s 10 vnitřními uzly

[1] http://wwwmagic.mppmu.mpg.de

Literatura

[1] R.K. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savický, S. Towers, A. Vaicilius, "Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope." *Nucl. Instr. Meth.*, A 516, pp. 511–528, 2004

[2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Belmont CA: Wadsworth, 1993

[3] J. Dvořák, P. Savický, "Softening Splits in Decision Trees Using Simulated Annealing", *Adaptive and Natural Computing Algorithms*, LNCS vol. 4431/2007, pp. 721–729, 2007

[4] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol. 27, pp. 861—874, 2006

[5] J.A. Nelder, R. Mead, "A simplex algorithm for function minimization.", *Computer Journal* vol. 7, pp. 308–313, 1965.

[6] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo — California, 1993