Selective Sampling (výběrové vzorkování)

Závěrečná zpráva projektu z P056
jaro 2002

Miroslav Doleček, <dolecek@fi.muni.cz>
Miroslav Křipač, <kripac@fi.muni.cz>

Abstrakt

Cílem projektu bylo implementovat předzpracovávací techniku výběrového vzorkování ve vylepšené verzi a otestovat ji na vhodných datech. Výběr klasifikátorů sestavujících výslednou množinu nebyl volen náhodně, ale na základě výsledků jejich učení nad náhodně vybranými iniciálními daty. Do výsledné množiny pak byly vybrány ty příklady z učících dat, na kterých takto zvolené klasifikátory dosáhly nejvyšší entropie.

Úvod

Výběrové vzorkování je jedna z předzpracovávacích metod strojového učení, která umožňuje z velkého množství učících dat vybrat pouze ta, která jsou pro proces učení relevantní. Velikost výsledné množiny pak může být výrazně menší při současném zachování přesnosti procesu učení. Celková doba potřebná k předzpracování a k učení pak bude výrazně nižší, než je doba učení na všech učících datech.

Relevantní data jsou při výběrovém vzorkování vybírána na základě výsledků klasifikátorů. Klasifikátor je obvykle velmi rychlý (i když ne tolik přesný) rozhodovací algoritmus, který je spuštěn na náhodně vybrané malé části učících dat. Počet klasifikátorů a velikost učících dat na kterých jsou naučeny je dáno parametrem výběrového vzorkování. Takto naučené klasifikátory se poté spustí na všechny příklady z učící množiny a zkoumá se, jak pro jednotlivý přiklad rozhodnou. Výsledná množina (jejíž velikost je v poměru k původní množině dána rovněž jako parametr) je pak sestavena na základě těchto rozhodnutí.

Popis implementace algoritmu

Tato implementace je napsaná v jazyce Perl a jako klasifikátor využívá lineární diskriminant LinDiscr, který je dostatečně rychlý. Jako cíl předzpracovaných dat jsme zvolili program Ltree sloužící ke tvorbě lineárních rozhodovacích stromů.

Pro realizaci algoritmu jsme použili dvě hlavní vylepšení. První z nich se týká výběru klasifikátorů. Program nejprve (z náhodně vybraných ne nutně disjunktních množin učících dat o velikosti zadané jako parametr na vstupu) sestaví dvojnásobné množství klasifikátoru, než je počet požadovaný pro rozhodovaní. Při tomto učení je jako výsledek, kromě výsledného rozhodovacího stromu, výstupem programu LinDiscr koeficient charakterizující výsledný model.

Na základě těchto koeficientu vybereme požadované množství klasifikátorů (polovinu) tak, aby se jejich koeficienty co nejvíce lišily. To způsobí, že rozhodovaní bude prováděno různými klasifikátory různě a umožní tak odhalit příklady, na které není jednoduché rozhodnout.

Druhé vylepšení spočívá ve výběru příkladů, kdy do výsledné množiny zařadíme prvky s nejvyšší entropií. Pro výpočet entropie se nakonec nejlépe osvědčila Shannonova metoda, která představuje míru různosti rozhodnutí klasifikátorů. Tedy příklad největší entropii má příklad, na kterém došlo k největšímu počtu rozdílných (špatných oproti správným) odpovědí. Do výsledné množiny pak zahrneme ty příklady, jejichž entropie dosáhne určitého prahu (zadaného jako parametr na vstupu).

Implementace výběru klasifikátorů

Po spuštění program nejprve načte potřebné parametry, případně doplní implicitně nastavené hodnoty. Poté po zjištění celkové velikosti učící množiny (a tím i velikosti množinu pro učení klasifikátorů) vygeneruje do pomocného pole pro každý vybíraný klasifikátor náhodná čísla v rozsahu velikosti vnitřní množiny. Tato uloží setříděně tak, že při načítání jednotlivých příkladů ze vstupní množiny učících dat sekvenčně pro každý příklad rozhodne, zda bude součástí daného klasifikátoru a uloží tato data do druhého pomocného pole. Poté pro každý klasifikátor uloží data z tohoto pole do jeho pomocného souboru a nad tímto souborem spustí program LinDiscr pro zjištění jednotlivých koeficientů. Seznam klasifikátorů indexovaný těmito koeficienty setřídí podle hodnoty koeficientu a vybere vždy sudé klasifikátory, čímž vznikne seznam klasifikátorů, které se navzájem nejvíce liší. Zároveň uložíme koeficienty takto vybraných klasifikátorů v binárním tvaru pro testovací běh programu LinDiscr.

Implementace výběru příkladů

Před zjištěním chování jednotlivých klasifikátorů program nejprve zjistí správné výsledky načtením všech testovacích příkladů. Poté pro každý příklad určí a uloží počet správných a špatných rozhodnutí klasifikátorů. Na základě toho stanoví entropii. V případě, že příkladů s entropii vyšší nebo rovnou než je stanovený parametr je méně než požadovaná výsledná množina, vybere algoritmus zbývající příklady náhodně ze zbylých příkladů (s nízkou entropií). Takto získané příklady jsou poté postupným průchodem učících dat zapsány do výsledné množiny.

Výsledky na ladících datech

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree	291s	13.95%(0.00%)	13.95%(0.00%)	**13.95%**(0.00%)
Ltree + SS 1	14s	15.50%(1.55%)	15.50%(1.55%)	15.50%(1.55%)
Ltree + **SS 2**	20s	14.94%(0.99%)	15.57%(1.62%)	**13.97%**(**0.02%**)
NBTree	?	14.10%(0.15%)	14.10%(0.15%)	14.10%(0.15%)
C4.5	?	15.54%(1.59%)	15.54%(1.59%)	15.54%(1.59%)

V tabulce jsou uvedeny záznamy výsledků testování několika různými metodami nad databází lidí (za úkol je rozhodnout ze zadaných údajů o dospělých lidech, zda jejich roční plat přesáhne 50 000 USD). Metoda Ltree je aplikace programu Ltree na celou učící množinu. Jedná se o nejdelší metodu, ale také nejpřesnější. Tato metoda bere v úvahu všechna učící data, proto dává vždy stejný výsledek. Metoda Ltree + SS 1 je spojení algoritmu Ltree s předzpracovávací metodou výběrového vzorkování v původní verzi, která rovněž nad stejnými daty dává vždy tutéž hodnoty. Doba trvání je v tomto případě součet doby potřebné k předzpracování a doby učení algoritmu Ltree nad předzpracovanou množinou. Metoda Ltree + SS 2 je námi zpracované řešení výběrového vzorkování. Testovali jsme dvanáctkrát po sobě nad týmiž daty a vlivem náhodného výběru klasifikátorů jsme dospěli k různým hodnotám. Má tedy smysl hovořit o průměrné, maximální a minimální chybě. Hodnota chyby udává vždy procento případů, ve kterých se daná metoda na testovacích datech zmýlila. Číslo v závorce je vždy zhoršení oproti nejlepší (a také nejdéle trvající) metodě Ltree. Výsledky metod NBTree a C4.5 jsme získali z přiložené dokumentace k datům. Jsou proto jen orientační a neobsahují údaj o čase.

Experimentální výsledky s různým nastavením výběrového vzorkování

Pro tento experimentální test jsme vybrali jiná data, než pro předchozí příklad. Jedná se o data reprezentující ručně psané číslice 0..9. Velikost učící množiny je 9494 záznamů. Velikost testovací množiny je 3948 záznamů. Cílem strojového učení je následně rozpoznat konkrétní číslice. Testovali jsme osmkrát po sobě. Následující tabulky prezentují výsledky s různým nastavením parametrů pro výběrové vzorkování.

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree bez SS	89.70s	5.94% (0.0%)	5.94% (0.0%)	5.94% (0.0%)

Počet klasifikátorů: 2
Iniciální část dat: 0.005 (37 záznamů)
Cílová část dat: 0.2 (1498 záznamů)
Entropie: 0.5

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree + **SS 2**	7.85s	16.61% (10.67%)	18.92% (12.98%)	14.00% (8.06%)

Počet klasifikátorů: 4
Iniciální část dat: 0.005 (37 záznamů)
Cílová část dat: 0.2 (1498 záznamů)
Entropie: 0.5

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree + **SS 2**	8.7s	13.12% (7.18%)	19.92% (13.98%)	8.00% (2.06%)

Počet klasifikátorů: 8
Iniciální část dat: 0.005 (37 záznamů)
Cílová část dat: 0.2 (1498 záznamů)
Entropie: 0.5

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree + **SS 2**	10.6s	12.17% (6.23%)	22.04% (16.1%)	9.51% (3.57%)

Počet klasifikátorů: 2
Iniciální část dat: 0.010 (74 záznamů)
Cílová část dat: 0.2 (1498 záznamů)
Entropie: 0.5

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree + **SS 2**	8.06s	16.93% (10.99%)	31.58% (25.64%)	10.77% (4.83%)

Počet klasifikátorů: 2
Iniciální část dat: 0.005 (37 záznamů)
Cílová část dat: 0.3 (2247 záznamů)
Entropie: 0.5

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree + **SS 2**	14.44s	13.61% (7.67%)	18.49% (12.55%)	9.00% (3.06%)

Počet klasifikátorů: 2
Iniciální část dat: 0.010 (74 záznamů)
Cílová část dat: 0.3 (2247 záznamů)
Entropie: 0.5

Metoda	Doba trvání	Prům. chyba	Max. chyba	Min. chyba
Ltree + **SS 2**	17.43s	10.64% (4.7%)	12.06% (6.12%)	8.14% (2.2%)

Domníváme se, že relativně větší odchylka chyby (narozdíl od učení bez výběrového vzorkování) je způsobena velkou variabilitou těchto dat. Tyto data jsou reprezentována 16 číselnými atributy, takže je možné velké množství jedinečných záznamů. Předchozí data byla reprezentována 15 atributy, z toho převážná většina atributů byla výčtového typu. Tím se proces učení velmi zjednodušuje.

Při testování jiných dat (databáze hřibů - jedlý/jedovatý), kde byly zastoupeny pouze výčtové atributy, byly průměrné chyby výrazně blízko průměrné chybě bez použití výběrového vzorkování, za velkého snížení časových nároků. Tyto data ovšem neobsahovala testovací část, proto jsme výsledky zde neuváděli. Učící data jsme rozdělili na 3/4 (nová učící množina) a 1/4 (nová testovací množina).

Závěr

Výše uvedenými vylepšeními se nám podařilo vytvořit algoritmus, který v nejlepším případě dosahuje téměř stejných výsledků, jako bez použití předzpracovávací metody za současného uchování nízké doby učení. Lze jej proto použít v případě velkého množství dat nebo pro případy, kdy je potřeba rozhodnout rychleji ale zároveň velice přesně.