BalDa

Vojtěch Havlíček

31.10.2012

BalDa (Balance Data) je nástroj pro vážení dat. Data určená pro modelování je vhodné před použitím v některých typech modelů ohodnotit váhami, tak aby méně časté záznamy s extrémějšími hodnotami, měly stejnou významnost jako časté záznamy s obvyklými hodnotami. Jedná se zejména o datově orientované modely a modely pracující s postupnými iteracemi vedoucími k optimálnímu řešení na základě chybových funkcí. Z hlediska typu dat je vážení vhodné provádět pro řady jejichž průběh je v čase víceméně vyrovnaný s krátkými úseky výrazné změny, což jsou např. řady průtoku, srážek apod.
Program BalDa je vytvořen v programovacím jazyku Fortran 90 a je určen pro operačni systém GNU/Linux. Program BalDa je volně dostupný na adrese

http://www.kvhem.cz/vyzkum/software/

Vážení dat je v programu BalDa možno provést pomoci třech algoritmů založených na metodě k-nejbližších sousedů (k-nearest neighbour – k-NN) a pomocí metody určující vzdálenost daného prvku od lineární nadroviny proložené k-nejbližšími sousedy. Uvedené metody jsou blíže popsany v (Vladislavleva et al., 2010).

Ovládání programu:

Po stažení je nutné program přeložit a sestavit. Provedení těchto úkonů je možné zadáním příkazu `make` v terminálu (v hlavním adresáři programu).
Vstupní data musí být ve formě textového souboru, ve kterém sloupce představují řady jednotlivých proměnných. Soubory se vstupními daty musí být umístěny ve složce "Inputs". Oddělovačem sloupců mohou být mezery nebo tabelátory. Data ve vstupním souboru nesmí být ukončena prázdnou řádkou a musí být bez hlavičky.
Nastavení programu pro výpočet je možné editací souboru "SETTINGS"v hlavním adresáři programu. V tomto souboru jsou následující možnosti nastavení:

Input File. Název souboru se vstupními daty.

Type of balancing. Metoda vážení dat. Možnosti jsou: *PRO* – provede vážení na základě vzdálenosti daného prvků od svých k nejbližších sousedů, *SUR* – provede vážení na

základě obklopení daného prvku svými k nejbližšími sousedy, *REM* – je kombinaci obou předcházejících metod, *NOL* – je vážení na základě odlehlosti prvku od lineární nadroviny procházející k nejbližšími sousedy, *ALL* – provede výpočet vah všemi výše uvedenými metodami.

No. of columns in file. Zadání počtu sloupců v souboru se vstupními daty.

No. of independent variable/s Počet nezávislých proměnných v souboru se vstupními daty, které budou použity pro výpočet vah.

Independent variable/s Určení nezávislých proměnných, které budou použity pro výpočet vah. Pro určení sloupců proměnných se zadává pořadí sloupce v souboru (bez tečky za hodnotou pořadí). Pořadí více sloupců proměnných jsou zadávány v řadě za sebou oddělené mezerou – *x1 x2 x3 ... xn*.

Dependent variable Pořadí sloupce, ve kterém je řada závislé proměnné.

Number of neighbours. Počet k nejbližších sousedů

Plot graphs Vytvoří grafický výstup pomoci R, je-li R nainstalováno. Možnosti jsou *Y* – vytvoření pdf souboru s grafem nebo *N* – tisk grafu do souboru nebude proveden. Grafický výstup je možný pouze pro výpočet s jednou nezávislou proměnnou a závislou proměnnou.

Výstupy programu jsou uloženy ve vytvořené složce "Outputs"(při novém spuštění programu je přemazána). Hlavním výstupem je textový soubor "*weights", ve kterém jsou v prvních sloupcích uloženy vybrané proměnné, které sloužily k výpočtu, a v posledním sloupci jsou vypočtené váhy.
V případě splnění podmínek pro vytvoření grafického výstupu, budou ve složce "Outputs"i pdf soubory s grafickým vyjádřením vah kombinací nezávislé a závislé proměnné.

Literatura:

Vladislavleva, E., Smits, G., den Hertog, D.: *On the importance of data balancing for symbolic regression.* IEEE Transactions on Evolutionary Computation 14(2), 252 –277 (2010)