Dolování dat pomocí Sammonova zobrazení

J.Doležal, J. Štěpanovský
Gymnázium, Špitálská 2;Gymnázium Třebíč
madrak@centrum.cz, stepanovsky@ic.cz

Abstrakt:

V tomto miniprojektu jsme se zabývali heuristikou a to konkrétně Harmony search. Tento postup jsme následně aplikovali v Sammonově zobrazení, jenž se využívá na pozměnění dimenzionality problému. Vytvořili jsme program, který to prováděl. Vyzkoušeli jsme dvě různé funkce na počítání Sammonova erroru. Na základních soustavách bodů jsme tyto dvě funkce porovnali.

1 Úvod

Sammonovo zobrazení je velmi užitečným nástrojem pro zpracování dat (tzv. data mining). Využívá se například v lékařství, human reseources, psychologii a mnoha technických odvětvích.

2 Náplň našeho miniprojektu

a. Účelové funkce

V našem programu využíváme více typů funkcí, které přiřazují našim sadám vygenerovaných bodů jejich míru "stresu" nebo-li správnosti. Tento stres se dá vypočítat více různými metodami Jedním z nich je jednoduché Sammonovo zobrazení.

$$E = \sum_{i>j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \qquad (1)$$

jsou zde původní vzdálenosti mezi původními body v původním prostoru, jsou vzdálenosti mezi nově vytvořenými body a *E* je zde stres. Jiná funkce se nazývá občas prodloužené Sammonovo zobrazení a využívá Bregmanových divergencí

$$E = \sum_{i>j} (d_{ij} ln \frac{d_{ij}}{d_{ij}^*} - d_{ij} + d_{ij}^*) \quad (2)$$

proměnné jsou zde úplně ty samé, jako v základním vzorci. Rozdíl mezi těmito dvěma vzorci je dle [2] v tom, že mají různé hodnoty stressu pro stejné rozdíly vzdáleností a zároveň je vzorec (2) lepší v tom, že zachovává více původní sousedy.

b. Heuristika-Harmony search

Harmony search je metoda, kterou jsme generovali nové body v novém N-rozměrném prostoru a tyto body měli být podobně vzdálené jako v původním prostoru.

Nejprve jsme vygenerovali 6 sad náhodných bodů v novém prostoru. Těmto sadám se přezdívá hudebníci. U každé ze sad jsme otestovali Sammonův error a to buď dle vzorce (1) nebo (2). Tyto errory jsme zapsali do jiné matice.

Vytvořili jsme novou sadu bodů (hudebníka), kterou jsme vytvořili pomocí následujícího postupu. Každý bod měl určitou pravděpodobnost, že se inspiruje v již

existující sadě bodů (nový hudebník imituje tón již existujícího). Pokud se tak stalo, bod mohl být ještě upraven (mohlo to zmutovat). Pokud se hudebník neinspiroval, bod vznikl naprosto náhodně v rámci mezí, které byly předem dány.

Novou sadu jsme opět otestovali tou samou funkcí jako sady předchozí. Poté jsme ji porovnali s nejhorší sadou (sada s nejvyšší errorem) a v případě, že se ukázala lepší, stávající sadu nahradila.

Proces nahrazování sad (muzikantů) novými měl konečný počet možných spuštění. Na tom obvykle závisela velikost erroru (přesnost).

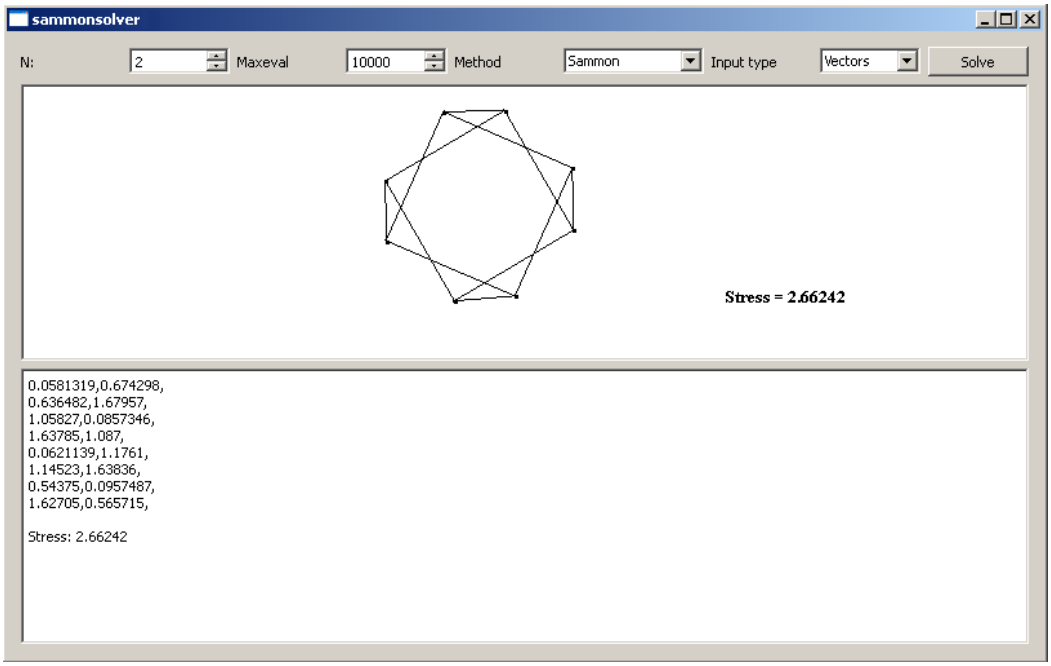
c. Samotný program

Program se skládá z grafického uživatelského rozhraní (GUI), které zajišťuje vkládání dat a parametrů, a z následujících metod:

Tabulka 1 Funkce

initHarmonySearch	Zinicializuje metodu zprostředkující Harmony Search
harmonySearch	Metoda zajištující harmony search (popsáno výše)
sammonReal	vlastní počítání erroru funkcí (1)
sammonExtendedReal	vlastní počítání erroru funkcí (2)
v2m	Převádí vektor na matici
m2d	Převádí matici na množinu vzdáleností mezi body
paint	Graficky znázorní výsledné body a vypíše jejich souřadnice
Main	Spustí ostatní funkce ve správném pořadí v závislosti na vstupních datech a parametrech

Příklad zobrazení krychle do roviny vytvořeným programem Sammonsolver pomocí funkce (1).



Obrázek 1 Sammonsolver

3 Shrnutí

Harmony search je metoda, která nachází lokální minimum a snaží se o nalezení globálního minima funkcí. Její síla a i zároveň slabina spočívá v náhodnosti výběru. Sammonovo mapování zobrazuje v multidimenzionálním prostoru do bodů v prostoru o menším počtu dimenzí a zprostředkovává přehlednější zobrazení dat. My jsme úspěšně spojili tyto dvě funkce do programu Sammonsolver a efektivně jsme je využili.

Poděkování

Děkujeme univerzitě ČVUT a FJFI za poskytnutí odborného vedení a literatury, bez nichž by tato práce nemohla nikdy vzniknout, a ostatním.

Reference:

[1]A Nonlinear Mapping for Data Structure analysis-John W. Sammon JR. *IEEE Transactions on Computers* **18,** str: 401–409

[2]Sun, J., Crowe, M., Fyfe, C., Extending metric multidimensional scaling with Bregman divergences, *Pattern recognition*, Elsevier, 2011, str. 1137.-1154.