ISSN 1648-8776
JAUNŲJŲ MOKSLININKŲ DARBAI. Nr. 1 (45). 2016 Doi: 10.21277/jmd.v1i45.45

EVOLIUCINIS POŽIŪRIS OPTIMIZAVIMO PROCESE

Agnė Dzidolikaitė
Vilniaus universiteto Matematikos ir informatikos institutas

Įvadas

Įvairiose žmogaus veiklos srityse nuolat susiduriama su daugiamačiais duomenimis. Tikriausiai nerastume tokios veiklos, kurioje apsieinama be daugiamačių duomenų. Vykstant techniniam progresui, tokių duomenų kiekis labai greitai didėja [1].

Medicinoje, technikoje ir kitose srityse tenka analizuoti daugiamačius duomenis. Šių duomenų apimtys vis didėja. Jiems nagrinėti ir kaupti pasitelkiama naujausia techninė ir programinė įranga.

Kaip jau minėta, duomenų apimtys auga didžiuliu greičiu, todėl tenka spręsti tam tikrus uždavinius, t. y. kaip tuos duomenis suprasti, komentuoti ir gauti reikiamą informaciją atsisakant nereikšmingų faktų. Dažniausiai kyla poreikis nuosekliau suvokti tokių duomenų sandarą: panašių objektų grupes (klasterius) ir ryškiai išsiskiriančius objektus (taškus atsiskyrėlius), objektų panašumą arba skirtingumą. Daugiamačius duomenis sunku suprasti, nes jie apibrėžia sudėtingą reiškinį ar objektą, turintį daug parametrų (rodiklių, ypatybių). Šie parametrai būna ne tik skaitiniai, bet ir loginiai, tekstiniai ar kitokie [2].

Kuo didesnės dimensijos duomenys yra, tuo sunkiau juos suvokti. Daugiamačiai duomenys pateikiami tokia forma, kad tyrėjui būtų lengviau suprasti duomenų sandarą, susidariusias grupes, tarpusavio ryšius ir pan. Toks duomenų pateikimo būdas galėtų būti vizualizavimas. Vizualizavimas – tai grafinis daugiamačių duomenų vaizdavimas. Grafinius duomenis žmogui lengviau suvokti. Daugiamačių duomenų vizualizavimas leidžia tirti ir atvaizduoti informaciją žmogui suprantamesne forma. Tyrėjas pats gali matyti duomenų grupavimosi tendencijas, nustatyti daugiamatės erdvės taškų tarpusavio artimumą ir priimti atitinkamus sprendimus.

Daugiamatės skalės

Įvairiuose moksluose paprastai sukaupiami dideli duomenų kiekiai. Žinoma daug įvairių metodų informacijai iš duomenų išgauti. Paprastai derinama statistinė analizė su tyrėjų patirtimi šiai užduočiai atlikti. Žmogaus euristiniai gebėjimai yra tinkami tada, kai matmenų skaičius neviršija trijų. Daugiamatės skalės leidžia duomenis, esančius didesnėje erdvėje, atvaizduoti į mažesnę erdvę. Daugiamatės skalės yra įtempimo funkcijos optimizavimas. Kadangi optimizavimo uždavinys yra multimodalus, turi būti taikomi globaliojo optimizavimo metodai.

Pradiniai daugiamačių skalių metodo duomenys yra kvadratinė simetrinė matrica, kurios elementai nusako analizuojamų objektų panašumą (skirtingumą). Paprasčiausiu atveju tai Euklido atstumų tarp objektų matrica. Tačiau bendruoju atveju tai nebūtinai turi būti atstumai griežtai matematine prasme [3].

Tarkime, kad kiekvieną n-matį vektorių $X_i \in R^n$, $i \in \{1,...,m\}$ atitinka mažesnio skaičiaus matmenų vektorius $Y_i \in R^d$, $d < n$. Atstumą tarp vektorių X_i ir X_j, vadinamą skirtingumu, pažymėkime δ_{ij}, o atstumą tarp vektorių Y_i ir Y_j – $d(Y_i, Y_j)$, $i,j = 1,..., m$. Atstumai tarp vektorių dažnai būna Minkovskio atstumai tarp taškų Y_i ir Y_j:

$$d_p(Y_i, Y_j) = \left(\sum_{k=1}^{m} |Y_{ik} - Y_{jk}|^p\right)^{1/p}$$

Kai $p = 2$, tada turime Euklido atstumus, o kai $p = 1$ – miesto kvartalo atstumus. Dėl matematinio patogumo atliekant įvairias procedūras dažniau naudojamas Euklido atstumas [4].

Naudojantis DS algoritmu bandoma atstumus $d(Y_i, Y_j)$ priartinti prie skirtingumų δ_{ij}. Jei naudojama kvadratinė paklaidos funkcija, E_{DS} gali būti užrašyta taip [5]:

$$E_{DS} = \sum_{i<j} w_{ij} \left(d(Y_i, Y_j) - \delta_{ij}\right)^2.$$

Paklaidos funkcija dar vadinama *Stress* (įtempimo) funkcija [6]. Tai yra santykinė paklaida, kurios formulė pateikiama žemiau:

$$E_{Stress} = \sqrt{\frac{\sum_{i<j} w_{ij} \left(d(Y_i, Y_j) - \delta_{ij}\right)^2}{\sum_{i<j} w_{ij}\, \delta_{ij}^2}}.$$

Genetiniai algoritmai

Genetiniai algoritmai mėgdžioja gyvąją gamtą [7]. Čia svarbiausia dalis yra kryžminimas. Tačiau algoritmo darbas pradedamas nuo inicializacijos, kai sukuriamas atsitiktinis individų sąrašas, kur kiekvie-

nas sąrašo elementas reiškia tam tikrą individą arba uždavinio sprendinį. Dažniausiai pradinė populiacija sudaroma iš šimtų ar tūkstančių atsitiktinių individų, kurių prisitaikymas yra prastas.

Vėliau sugeneruota populiacija įvertinama pagal kiekvieno nario prisitaikymą. Individo prisitaikymą galima skaičiuoti įvairiai. Tarkime, kad duotasis individas koduoja lygties sprendinį. Vadinasi, kuo individo pateiktas sprendinys artimesnis lygties sprendiniui, tuo didesnė jo prisitaikymo reikšmė, ir atvirkščiai.

Daugiamačių skalių atveju genais laikome objektus vaizduojančius taškus. Kryžminant dalis taškų paimama iš vieno ir dalis iš kito individo [8].

Mutacijos metu atsitiktinai pakeičiama viena ar daugiau geno komponenčių ir gaunama nauja individo savybė [9]. Galima geno komponentes vaizduoti nulių ir vienetų sekomis. Tada mutacijos atveju nulis keičiamas vienetu, o vienetas – nuliu [10].

Daugiamačių skalių algoritmas šiame darbe atliktas jungiant jį su genetiniu algoritmu. Tokio algoritmo pseudokodas būtų toks:

1: *Atsitiktinai generuoti individų populiaciją.*
2: *Apskaičiuoti kiekvieno individo mažiausiųjų kvadratų įtempimo funkciją.*
3: ***for*** *o kartų* ***do****.*
4: ***for*** *populiacijos dydžiui (l)* ***do****.*
5: *Atsitiktinai pasirinkti du populiacijos individus.*
6: *Kryžminti šiuos du individus (tėvus) atsitiktinai paimant genų iš kiekvienos tėvų chromosomos ir sukeičiant tuos genus tam tikroje chromosomos vietoje.*
7: *Su tam tikra mutacijos tikimybe atlikti naujai gauto palikuonio mutacijas.*
9: ***if*** *naujasis palikuonis turi geresnę prisitaikymo funkcijos reikšmę nei populiacijos individas su prasčiausia prisitaikymo funkcijos reikšme; tada pakeičiame pastarąjį individą.*
10: ***end do****.*
11: ***end do****.*

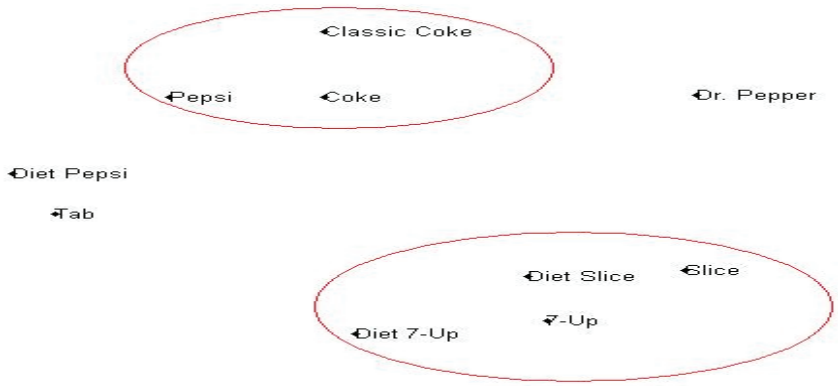
Eksperimentinis tyrimas

Čia tiriame dvi duomenų aibes (gaiviųjų gėrimų ir farmakologinių duomenų) [1]. Pirmoji duomenų aibė naudojama analizuojant eksperimentą, kuris buvo atliktas su 38 studentais. Studentai ragavo 10 skirtingų gaiviųjų gėrimų ir turėjo pasakyti, kaip skiriasi kiekvienos gaiviųjų gėrimų poros skonis skalėje nuo 1 iki 9. 1 reiškia, kad gaivieji gėrimai yra labai panašūs, o 9 reiškia, kad tos gaiviųjų gėrimų poros skoniai yra visiškai skirtingi. Daugiamačių duomenų aibė susideda iš gaiviųjų gėrimų skirtingumų sumų.

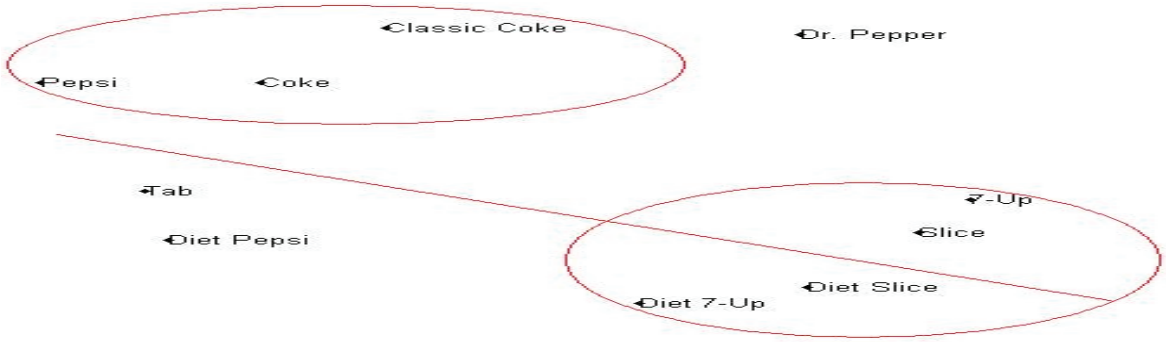
Antroji duomenų aibė rodo, kaip stipriai ligandai jungiasi prie baltymų. Ligandai yra mažos molekulės, kurios jungiasi prie baltymų ir veikia kaip natūralūs neurotransmiteriai ar vaistai, aktyvuodami ir blokuodami baltymus.

Daugiamačių skalių algoritmas užrašytas *Java* programavimo kalba. Santykinės paklaidos minimizavimas atliktas taikant genetinį algoritmą daugiamatėms skalėms. Algoritmo darbas sustabdytas po 1 000 kartų. Populiacija susideda iš 100 individų. Eksperimentas su kiekviena duomenų aibe pakartotas 30 kartų. Buvo sukurta 30 vaizdų. Tada atrinkta po du vaizdus kiekvienos duomenų aibės rezultatams analizuoti. Vaizdai parinkti taip, kad galėtume stebėti tam tikras duomenų aibių savybes. Pirmojo vaizdo duomenų aibė turi mažiausią santykinę paklaidą, o antrojo vaizdo duomenų aibės santykinės paklaidos reikšmė yra didesnė.

Pirmajame paveiksle matome du vaizdus, atitinkančius eksperimentą su gaiviaisiais gėrimais. Po kiekvienu vaizdu pateikiama santykinės paklaidos reikšmė.



a) santykinė paklaida yra lygi 0,201952



b) santykinė paklaida yra lygi 0,203886

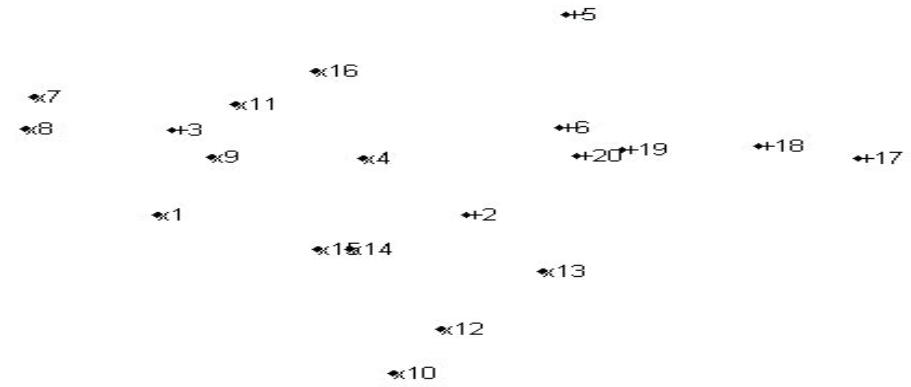
1 pav. Dviejų grupių gaiviųjų gėrimų vaizdas

Iš dešimt skirtingų gaiviųjų gėrimų vaizdų matome, kad kai kurie gaivieji gėrimai sudaro klasterius. *Diet 7-Up*, *Diet Slice*, *7-Up* ir *Slice* sudaro vieną klasterį, kitas klasteris susidaro iš *Pepsi, Coke* ir *Classic Coke.* Kartais galime nubrėžti liniją tarp dietinių gėrimų ir nedietinių gėrimų (žr. 1 pav. b). Santykinės paklaidos reikšmė (žr. 1 pav. a) yra mažesnė nei santykinės paklaidos reikšmė (žr. 1 pav. b), tačiau dietiniai gėrimai yra atskiriami linija nuo nedietinių gėrimų (žr. 1 pav. b) nepaisant to, kad šio vaizdo duomenų aibės santykinės paklaidos reikšmė yra didesnė.

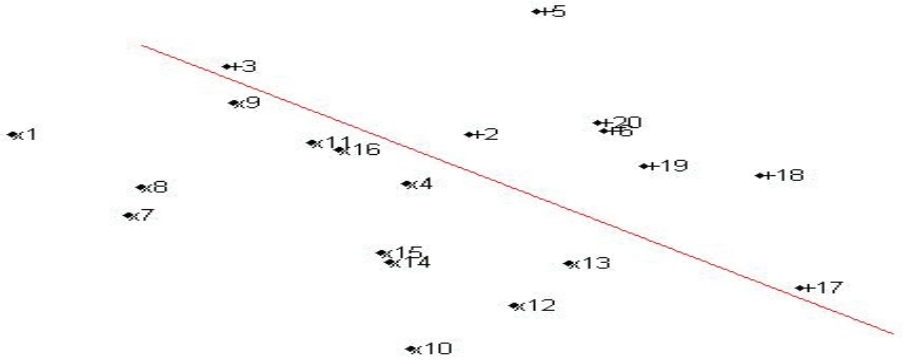
Du vaizdai atitinka eksperimentą su farmakologiniais duomenimis (žr. 2 pav.). Po kiekvienu vaizdu pateikiama santykinės paklaidos reikšmė [11, 12]. Čia atvaizduoti ligandai, pritaikius genetinį algoritmą daugiamatėms skalėms. Kaip buvo minėta anksčiau, ligandai yra mažos molekulės, kurios jungiasi prie baltymų ir keičia jų funkcijas aktyvuodamos ar blokuodamos juos. Paveiksluose (žr. 2 pav. a–b) aktyvuojantys ligandai vaizduojami '+', o blokuojantys ligandai vaizduojami 'x'. Paveiksle (žr. 2 pav. b) aktyvuojančius ligandus galima atskirti linija nuo blokuojančių ligandų. Santykinės paklaidos reikšmė paveiksle (žr. 2 pav. a) yra mažesnė nei paveiksle (žr. 2 pav. b), tačiau aktyvuojantys ligandai nuo blokuojančių ligandų yra atskiriami linija (žr. 2 pav. b) nepaisant to, kad šio vaizdo duomenų aibės santykinė paklaida yra didesnė. Toks pat rezultatas buvo pastebėtas ir aptartas analizuojant gaiviųjų gėrimų aibės vaizdus.

Su farmakologinių ir gaiviųjų gėrimų duomenimis atlikta ir kitų eksperimentų. Pavyzdžiui, duomenys tirti dirbtiniais neuroniniais tinklais ir stebėta, kaip susidaro duomenų klasteriai. Mokslinėje literatūroje siūloma jungti pagrindinių komponenčių ir evoliucinį metodus. Šiuo nauju metodu gauti duomenys atvaizduoti grafiškai ir gauti tam tikri vaizdai, gana panašūs į straipsnio autorės gautus vaizdus. Analizuojami ir stiklų bei gaiviųjų gėrimų duomenys, taikant dirbtinį neuroninį tinklą, sujungtą su genetiniu algoritmu. Neuroninis tinklas lengvai „įstringa" lokaliajame minimume, todėl, sujungus dirbtinį neuroninį tinklą su genetiniu algoritmu, naujai gauto hibridinio algoritmo efektyvumas yra didesnis nei minėtų dviejų algoritmų atskirai. Mokslininkai yra aprašę hibridinį neuroninį-genetinį algoritmą, kuris net su 94,44 proc. tikimybe suformuoja aiškius duomenų klasterius.

Mokslininkai yra taikę ir daugiau algoritmų, norėdami atvaizduoti farmakologinius ir gaiviųjų gėrimų duomenis. Straipsnio autorės algoritmas pateikia vaizdus su duomenų klasteriais. Šis metodas nuo kitų literatūroje aprašytų metodų skiriasi tuo, kad tiria ne tik optimalių sprendinių vaizdus, bet ir neoptimalių sprendinių vaizdus. Šiuo darbu siekta parodyti, kad ne tik optimalių sprendinių vaizdai gali duoti reikšmingų rezultatų. Reikšmingais rezultatais laikome tokius vaizdus, kuriuose galime stebėti aiškias atskiras duomenų grupes (klasterius).



a) santykinė paklaida yra lygi 0,100078



b) santykinė paklaida yra lygi 0,240342

2 pav. Dviejų ligandų grupių vaizdai

Išvados

1. Genetiniu algoritmu daugiamatėms skalėms gauti vaizdai gali duoti gerų rezultatų net ir tada, kai nėra globaliojo optimumo. Verta atvaizduoti keletą vaizdų ir parinkti iš jų geriausią. Geriausias vaizdas gali atitikti lokalųjį optimumą. Nepaisant to, šio vaizdo nederėtų atmesti kaip nereikšmingo.
2. Euristinius algoritmus, iš kurių straipsnyje nagrinėjami genetiniai algoritmai, galima taikyti daugiamatėms skalėms sprendžiant globaliojo optimizavimo uždavinius.

Literatūra

1. Dzemyda G., Kurasova O., Žilinskas J., 2008, *Daugiamačių duomenų vizualizavimo metodai*. Vilnius: Mokslo aidai.
2. Karbauskaitė R., 2005, Daugiamačių duomenų vizualizavimo metodų, išlaikančių lokalią struktūrą, analizė. *Daktaro disertacija*. Vytauto Didžiojo universitetas, Matematikos ir informatikos institutas.
3. Groenen P., Velden M., 2004, *Multidimensional Scaling*. Econometric Institute Report EI 2004-15.
4. Žilinskas A., Žilinskas J., 2006, On Multidimensional Scaling With Euclide and City Block Metrics. *Ūkio ir technologijos vystymas*. Vol. XII. Nr. 1. P. 69–75.
5. Leeuw J., 1977, *Applications of Convex Analysis to Multidimensional Scaling*. Holland.
6. Miyano H., Inukai J., 1982, Sequential Estimation in Multidimensional Scaling. *Psychometrika*. 47. P. 321–361.
7. Dzemyda G., Šaltenis V., Tiešis V., 2007, *Optimizavimo metodai*. Vilnius: Mokslo aidai.
8. Torn A., Žilinskas A., 1989, *Lecture Notes in Computer Science*. Vol. 350, Springer Verlag, Berlin.
9. Žilinskas A., 2005, *Matematinis programavimas*. Kaunas: VDU leidykla.
10. Enrique A., Dorronsoro B., 2008, *Cellular Genetic Algorithms*. Springer.
11. Lančinskas A., Ortigosa P. M., Žilinskas J., 2015, Parallel optimization algorithm for competitive facility location. *Mathematical Modelling and Analysis*. 20 (5). P. 619–640.
12. Žilinskas J., Goldengorin B., Pardalos P. M., 2015, Pareto-optimal front of cell formation problem in group technology. *Journal of Global Optimization*. 61 (1). P. 91–108.

Summary

EVOLUTIONARY APPROACH TO OPTIMIZATION

A. Dzidolikaitė

Large amounts of data are used in many science fields. Such data are often visualised to better understand them. There are many multidimensional data visualization methods. One of the best known methods among them is multidimensional scaling. Using this method high dimensional data are visualized into a two or three dimensional space. It is showed in the paper that multidimensional scaling can be used with genetic algorithm to solve optimization problems.

Keywords: multidimensional scaling, evolutionary algorithms, optimization.

Santrauka

EVOLIUCINIS POŽIŪRIS OPTIMIZAVIMO PROCESE

A. Dzidolikaitė

Daugelyje mokslo šakų naudojami dideli duomenų kiekiai. Kad tokie duomenys būtų geriau suprantami, jie dažnai yra vizualizuojami. Egzistuoja daugybė daugiamačių duomenų vizualizavimo metodų. Daugiamatės skalės yra vienas iš žinomiausių metodų. Taikant daugiamates skales, didesnių matmenų duomenys atvaizduojami į dviejų ar trijų matmenų erdvę. Šiame straipsnyje parodyta, kad daugiamatės skalės gali būti naudojamos kartu su genetiniu algoritmu optimizavimo uždaviniams spręsti.

Prasminiai žodžiai: daugiamatės skalės, evoliuciniai algoritmai, optimizavimas.

Įteikta 2016-01-15
Priimta 2016-06-15