Ataskaitinė informatikos krypties doktorantų konferencija
2017-10-17

Ataskaita už 2016 – 2017 mokslo metus

Doktorantė: Jelena Liutvinavičienė

Darbo vadovė: Prof. dr. Olga Kurasova

Bendra informacija

- **Disertacijos pavadinimas:** „Didelės apimties duomenų vizuali analizė“
- **Darbo vadovė:** doc. dr. Olga Kurasova
- **Doktorantūros pradžia:** 2014 m.
- **Planuojama doktorantūros pabaiga:** 2018 m.

Informacija apie tyrimą

- **Tyrimo objektas:**
  - Didelės apimties duomenys
- **Tyrimo tikslas:**
  - Pasiūlyti metodologiją, kuri leistų vizualizuoti didelės apimties duomenis integruojant kelis dimensijų mažinimo metodus, siekiant iš vaizdo gauti kuo daugiau vertingos informacijos apie vizualizuojamus duomenis.

Informacija apie tyrimą

- **Tyrimo uždaviniai:**
  - Analitiškai apžvelgti ir palyginti didelės apimties duomenų vizualizavimo metodus, juos įgyvendinančius įrankius bei technologijas, įgalinančias analizuoti didelės apimties duomenis.
  - Išnagrinėti vizualizavimo metodų, pagrįstų dimensijų mažinimu, tikslumo įvertinimo matus.
  - Pasiūlyti metodologiją, kuri leistų vizualizuoti didelės apimties duomenis.
  - Sudaryti programų sistemos prototipą, kuriame būtų pasiūlyta didelės apimties duomenų vizualizavimo metodologija.
- **Planuojami rezultatai:**
  - Didelės apimties duomenų vizualizavimo metodologija ir programų sistemos prototipas.

2016–2017 m. m. darbo planas

- **Mokslinių tyrimų planas:**
  - Naujo didelės apimties duomenų vizualizavimo būdo, kuris leistų vizualizuoti duomenis keliomis pakopomis, sukūrimas.
  - Sukurto vizualizavimo būdo eksperimentinis tyrimas analizuojant keletą testinių didelės apimties duomenų aibių.
  - Praktinių sričių, kuriose būtų tinkamas sukurtas vizualizavimo būdas, nustatymas bei jo pritaikymas praktiniam uždaviniui.

2016–2017 m. m. darbo planas

- **Rezultatų pristatymo planas:**
  - Dalyvavimas Lietuvos kompiuterininkų sąjungos organizuojamame renginyje „Kompiuterininkų dienos 2017“, vyksiančiame 2017 m. rugsėjo mėn.
  - Dalyvavimas 25-oje tarptautinėje konferencijoje „Computer Graphics, Visualization and Computer Vision 2017 (WSCG 2017)“, vyksiančioje 2017 m. gegužės 29 d. – birželio 2 d. Pilzene, Čekijos Respublikoje.
- **Mokslinių publikacijų planas:**
  - Planuojamas mokslinis straipsnis *Informacijos mokslų* žurnale.
  - Planuojamas mokslinis straipsnis *Baltic Journal of Modern Computing* žurnale.

Ataskaita už 2016–2017 m. m.

- **Dalyvavimas konferencijose 2016–2017 m. m.:**
  - 2016-10-13 – 2016-10-15 dalyvauta tarptautinėje konferencijoje ***„22nd International Conference on Information and Software Technologies“*** (ICIST 2016), pranešimas ***„Parallel computing for dimensionality reduction“***,
  - 2016-12-01 – 2016-12-03 7-oje mokslinėje konferencijoje ***„Duomenų analizės metodai programų sistemoms“***, Druskininkuose, pristatytas stendinis pranešimas ***„Multi-level Method for Big Data Visualization“***,
  - 2017-09-21 – 2017-09-22 Lietuvos kompiuterininkų sąjungos organizuotame renginyje ***„Kompiuterininkų dienos – 2017“***, vykusioje XVIII kompiuterininkų konferencijoje, Kaune, skaitytas pranešimas ***„Daugiamatiškumo mažinimo metodai: greičio ir tikslumo palyginimas“***.

Ataskaita už 2016–2017 m. m.

- **Publikacijos 2014–2015 m. m.:**
  - Zubova J., Kurasova O. (2014). ***Challenges of big data visualization***. *6th International Workshop on Data Analysis Methods for Software Systems [abstracts book]*, Druskininkai, Lithuania, December 4-6, 2014. ISBN 9789986680505. p. 59.
  - Zubova J., Kurasova O. (2015). ***Didelių duomenų vizualizavimo metodai ir įrankiai***. *Informacijos mokslai 73*: 113–126. Vilnius: Vilniaus universiteto leidykla. ISSN 1392-0561.

Ataskaita už 2016–2017 m. m.

- **Publikacijos 2015–2016 m. m.:**
  - Zubova J., Kurasova O., Medvedev V. (2015). ***Visual Analytics for Big Data****. 7th International Workshop on Data Analysis Methods for Software Systems [abstracts book]*, Druskininkai, Lithuania, December 3-5, 2015. ISBN: 9789986680581. p. 53-54.
  - Liutvinavičius M., Zubova J., Sakalauskas V. (2016). ***Financial crisis prediction: behavioural finance approach for stock market forecasting.*** *Fourth International Symposium in Computational Economics and Finance [Symposium Proceedings]*, Paris, France, April 14-16, 2016.
  - Liutvinavičius M., Zubova J., Sakalauskas V. (2016). ***Finansų rinkų prognozavimas remiantis investuotojų nuotaikų indikatoriumi.*** *Informacinės technologijos*, 2016 m. balandžio 28 d., Kaunas, Lietuva: Vytauto Didžiojo universitetas. ISSN 2029-249X. p. 17-22.

Ataskaita už 2016–2017 m. m.

- **Publikacijos 2016–2017 m. m.:**
  - Zubova J., Kurasova O. (2014). Multi-level method for big data visualization. *8th International Workshop on Data Analysis Methods for Software Systems [abstracts book]*, Druskininkai, Lithuania, December 1-3, 2016. ISBN 9789986680611. p. 70.
  - Zubova J., Liutvinavičius M., Kurasova O. (2016). Parallel Computing for Dimensionality Reduction. *Information and Software Technologies. Proceedings of 22nd International Conference, ICIST 2016. Communications in Computer and Information Science (639)*, Springer, ISBN: 9783319462530. p. 230-241.
  - Zubova J., Kurasova O., Liutvinavičius M. (2017). Dimensionality reduction for financial data visualization. *Informacinė visuomenė ir universitetinės studijos (IVUS 2017).* Konferencijos pranešimų medžiaga. ISSN 2029-249X.
  - Liutvinavicius M., Zubova J., Sakalauskas V. Behavioural Economics Approach: Using Investors Sentiment Indicator for Financial Markets Forecasting. *Baltic Journal of Modern Computing, Vol. 5 (2017), No. 3, 275-294.*
  - Zubova J., Kurasova O., Liutvinavičius M. Dimensionality reduction methods: the comparison of speed and accuracy. *Information Technology and Control [In press],* (žurnalas turi cituojamumo rodiklį *Clarivate Analytics Web of Science* duomenų bazėje, IF 2016: 0.475).

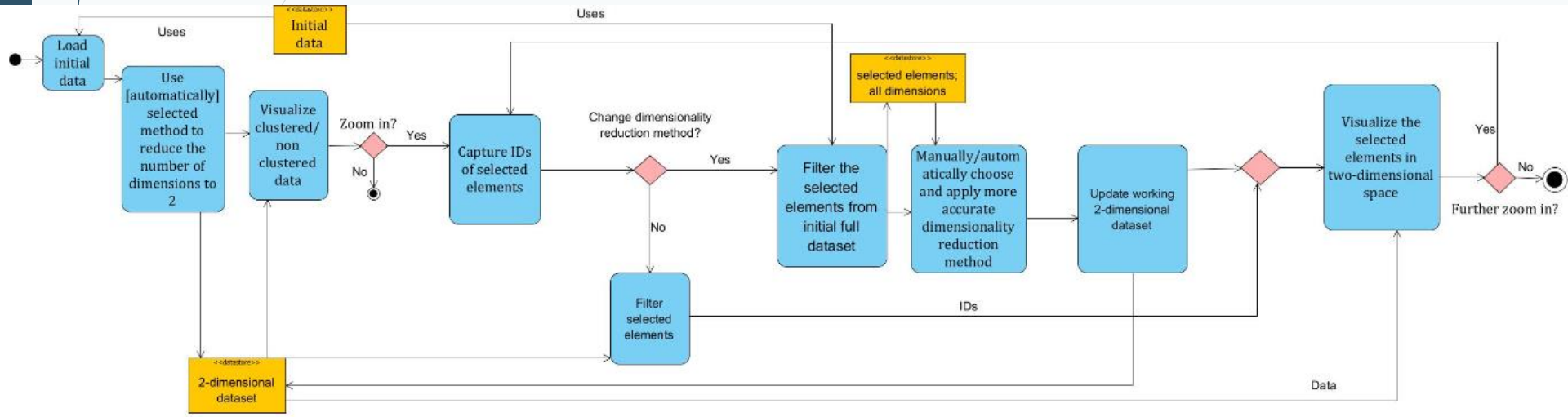
Ataskaita už 2016–2017 m. m.

- **2016–2017 m. m. gauti moksliniai rezultatai:**
  - Pasiūlytos didelės apimties duomenų vizualizavimo metodologijos pagrindinės sudedamosios dalys.
  - Išnagrinėti vizualizavimo metodų, pagrįstų dimensijų mažinimu, tikslumo įvertinimo matai. Dimensijų mažinimo metodų tikslumui įvertinti tyrimuose naudoti trys matai: Stress funkcijos reikšmė, Spirmano (Spearman) koreliacijos koeficientas ir Shannon entropijos koeficientas. Buvo matuojamas metodų vykdymo laikas.
  - Eksperimentiniai tyrimai atlikti su testiniais duomenimis: atsitiktinai sugeneruotais neklasterizuotais, atsitiktinai sugeneruotais klasterizuotais ir realiais finansiniais duomenimis. Buvo nustatyta, kad duomenų pobūdis neturi ženklios įtakos jų apdorojimo greičiui, tačiau tai turi įtakos dimensijų mažinimo tikslumui. Klasterizuotus duomenis galima tiksliau atvaizduoti dvimatėje erdvėje negu neklasterizuotus. Geriausi tikslumo rodikliai buvo pasiekti apdorojant realius duomenis.

Siūloma metodologija

- Duomenų vizualizavimas yra paremtas dimensijų mažinimo metodais.
- Metodologija remiasi principu, jog visas procesas yra skaidomas į atskirus etapus.
- Kiekviename etape konkretus metodas yra parenkamas priklausomai nuo duomenų kiekio ir jų pobūdžio.
- Teorinė mokslinių darbų apžvalga leidžia daryti prielaidą, jog vieni metodai yra greitesni, tačiau mažiau tikslūs, o kiti atvirkščiai – tikslesni, tačiau lėtesni.

Metodologijos algoritmo schema



- Tikslas – sukurti sistemos prototipą, realizuojantį siūlomą metodologiją.

Analizuoti dimensijų mažinimo metodai

- Multidimensional Scaling (MDS)
- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Principal Curve
- Locally Linear Embedding (LLE)
- Isometric Mapping (Isomap)

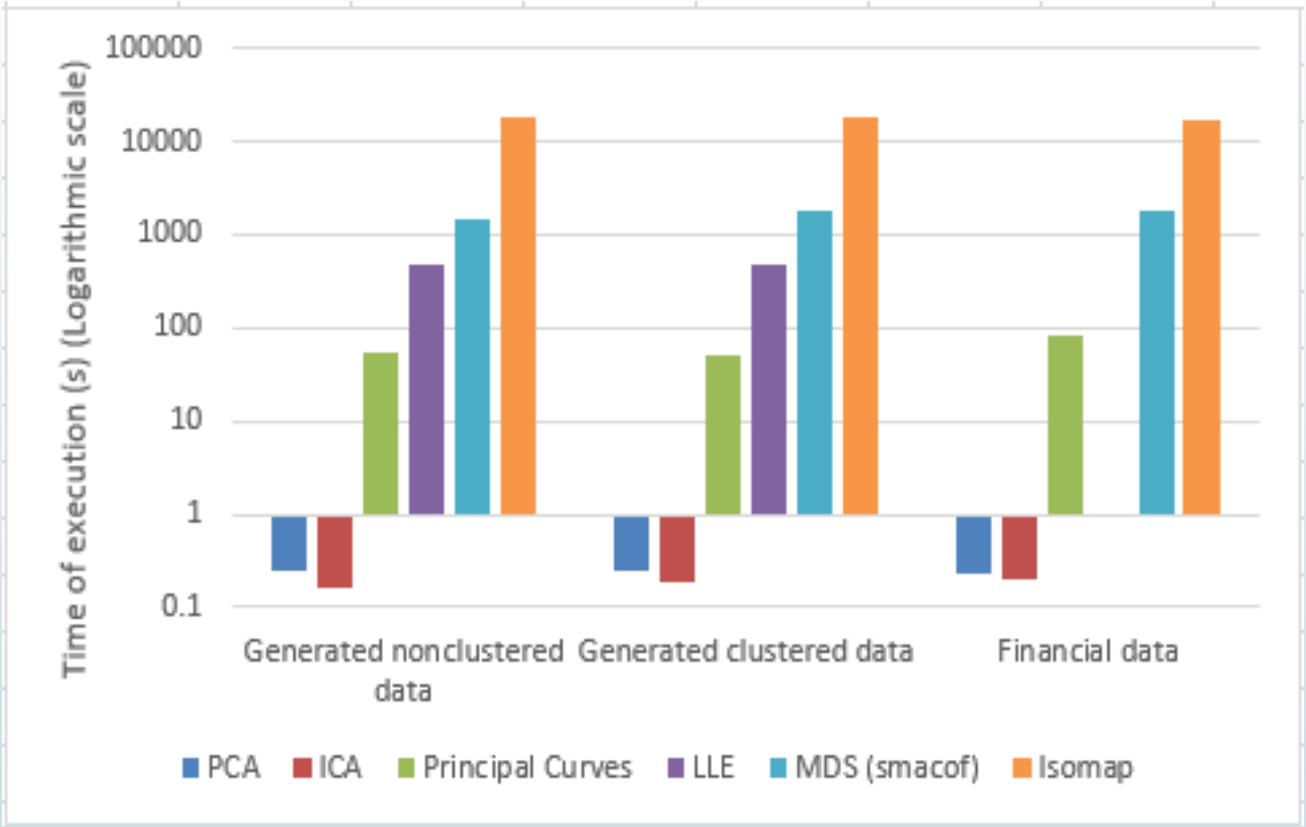
Testiniai duomenys

- Atsitiktinai sugeneruoti neklasterizuoti duomenys
- Atsitiktinai sugeneruoti klastesrizuoti duomenys
- Realūs finansiniai duomenys

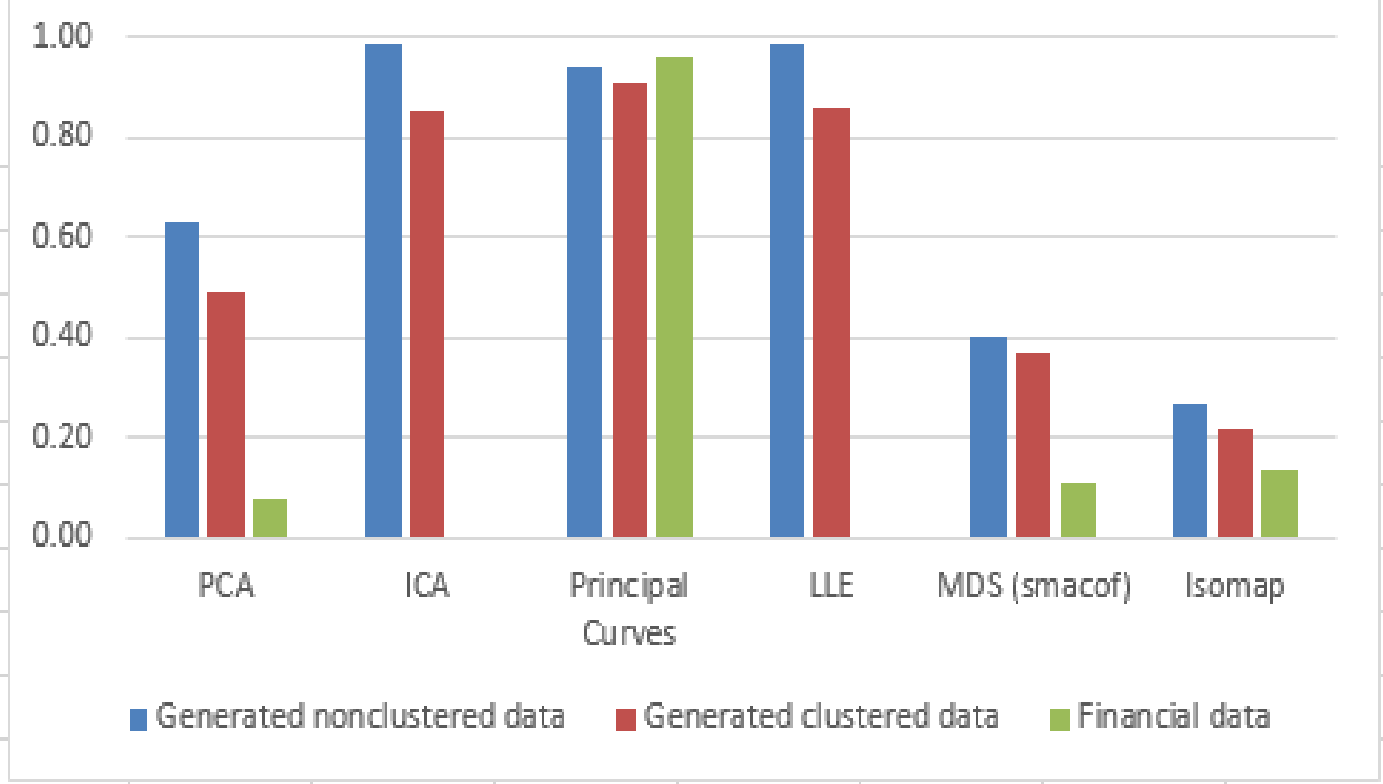
Metodų palyginimas

- Greitis – vykdymo laikas
- Tikslumas:
  - Stress
    - Daugiamačių skalių metodų kvadratinė paklaidos funkcija.
  - Spirmeno (Spearman) koreliacijos koeficientas
    - Ranginis kriterijus, kuris ryšio stiprumui įvertinti naudoja ne pačias kintamųjų reikšmes, o jų rangus. Galimos reikšmės nuo -1 iki 1.
  - Shannon entropijos koeficientas
    - Kriterijus, parodantis, kaip tiksliai tam tikru metodu gauta duomenų projekcija išlaiko informacijos kiekį, kurį turėjo pradinė duomenų aibė.

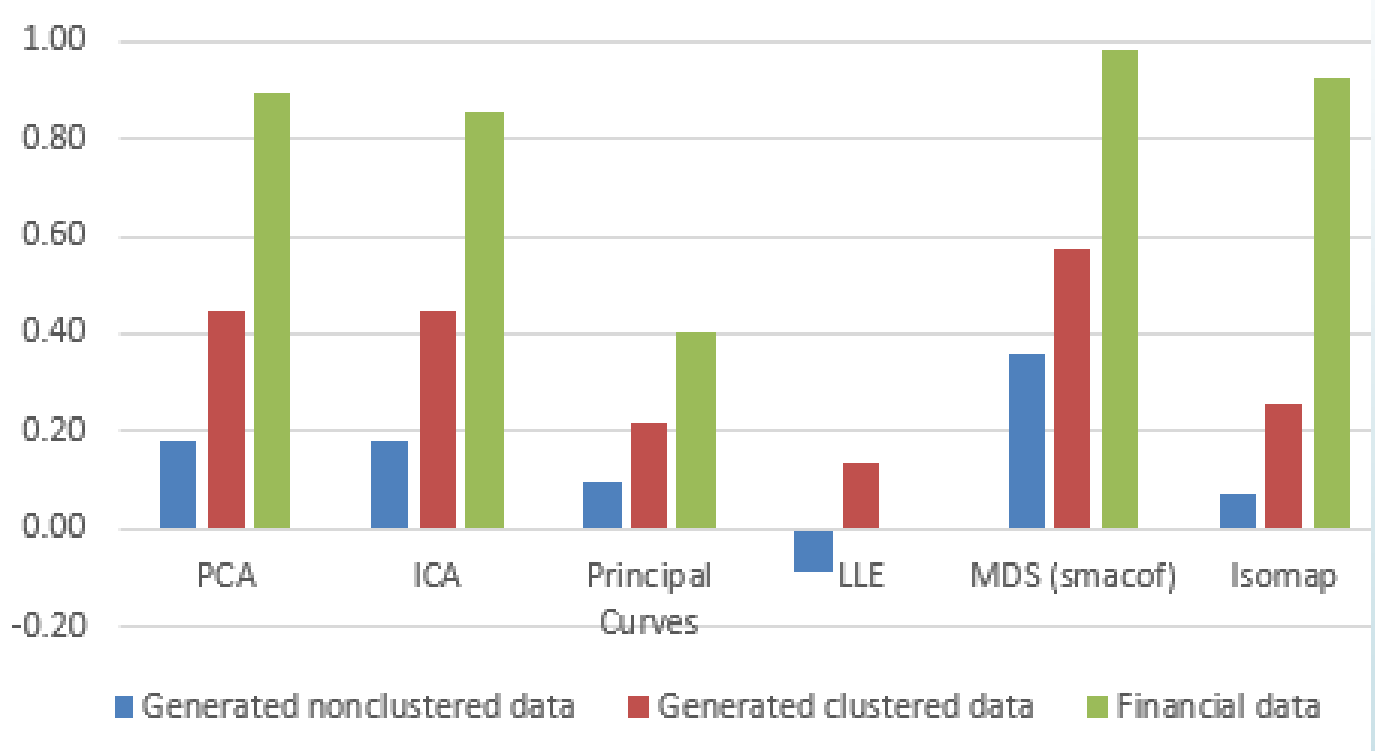
Skirtingo pobūdžio duomenų vykdymo laikai



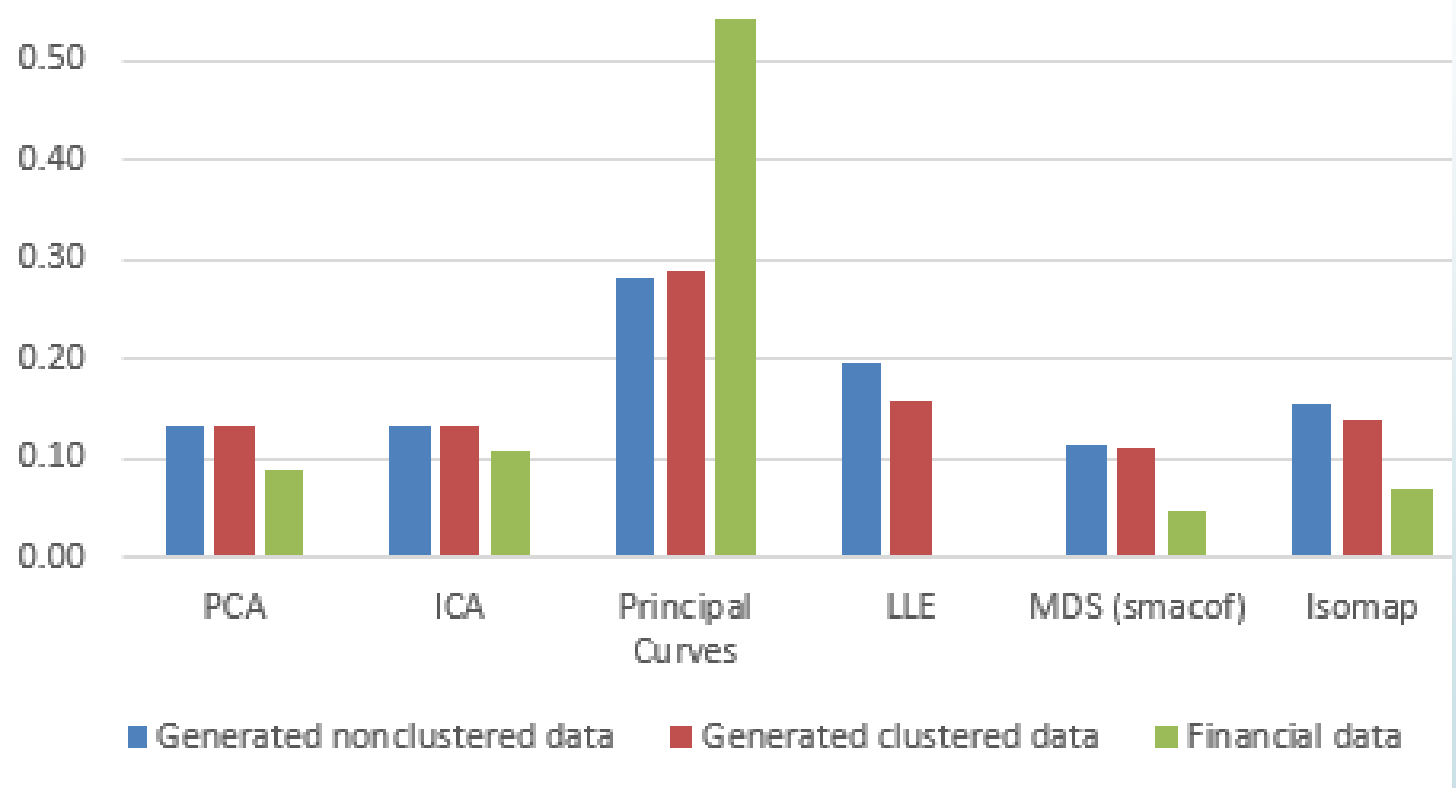
Tikslumo palyginimas: Stress



Tikslumo palyginimas: Spearman‘o koeficientas



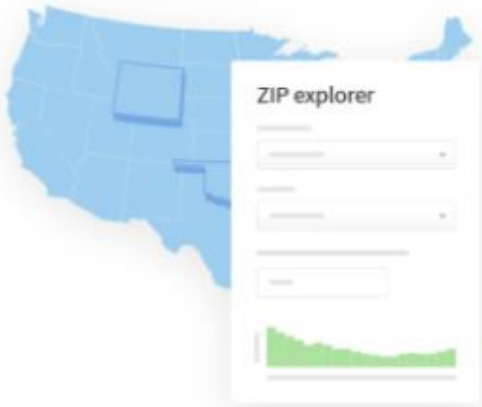
Tikslumo palyginimas: Shannon‘o entropija



RStudio®

RStudio

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.



Shiny

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

R Packages

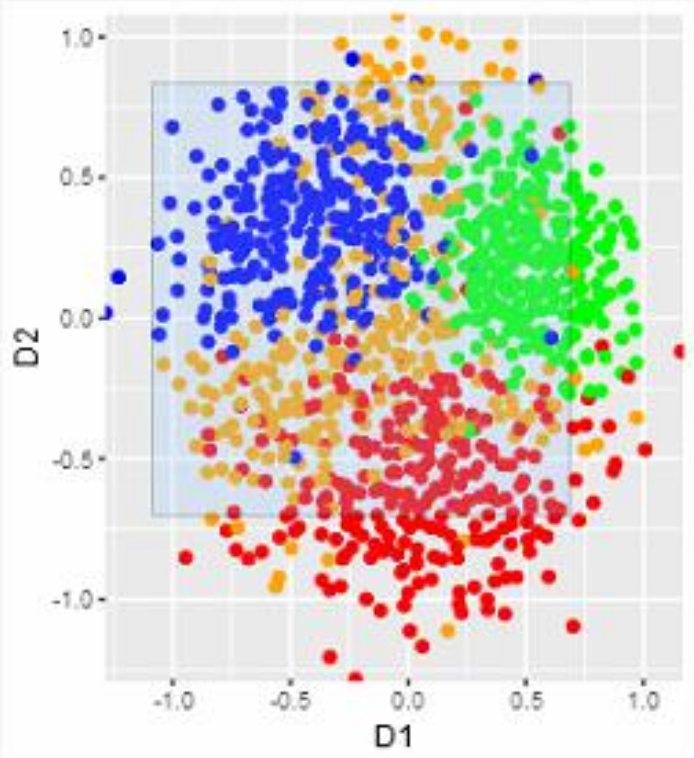
Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

Išvados

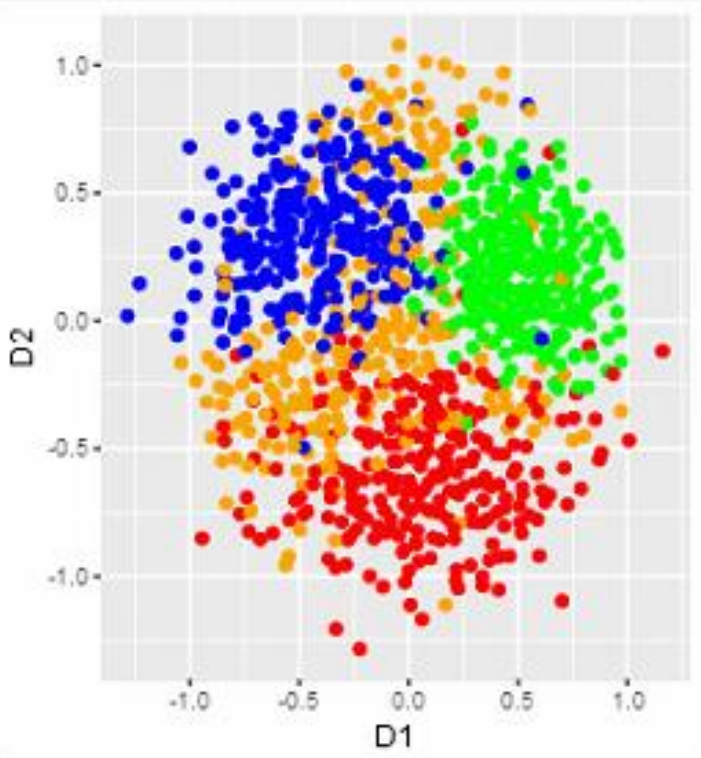
- Šis tyrimas skirtas didelių duomenų vizualizavimui pritaikant dimensijų mažinimo metodus. Kiekviename etape konkretus metodas parenkamas atsižvelgiant į jo greitį ir tikslumą.
- Buvo atliktas metodų greičio ir tikslumo palyginimas. Tam panaudoti trijų rūšių duomenys: atsitiktinai sugeneruoti neklasterizuoti, atsitiktinai sugeneruoti klasterizuoti ir realūs finansiniai duomenys.
- Atsitiktinai sugeneruotų duomenų atveju buvo patvirtintos šios taisyklės: Didesnis objektų kiekis lemia ilgesnį vykdymo laiką. Tuo tarpu pradinis dimensijų kiekis vykdymo laikui įtakos neturi (mažinant dimensijų kiekį iki dviejų). Tikslumo atžvilgiu situacija yra priešinga: objektų kiekis neturi įtakos dimensijų mažinimo tikslumui, didesnis pradinis dimensijų kiekis lemia mažesnį duomenų atvaizdavimo tikslumą (mažinant dimensijų kiekį iki dviejų).
- Tyrinėjant realius duomenis, pastebėta, jog didesnis objektų kiekis vis dėlto gali lemti ilgesnį duomenų apdorojimo laiką. Dalis metodų nesugebėjo apdoroti realių duomenų. Šiuo atveju taip pat pastebėta, jog didesnis objektų kiekis gali užtikrinti tikslesnį jų atvaizdavimą dvimatėje erdvėje.
- Bendrąja prasme duomenų pobūdis neturi ženklios įtakos jų apdorojimo greičiui. Tačiau tai turi įtakos dimensijų mažinimo tikslumui. Klasterizuotus duomenis galima tiksliau atvaizduoti dvimatėje erdvėje negu neklasterizuotus. Geriausi tikslumo rodikliai buvo pasiekti apdorojant realius duomenis.

Siūlomos metodologijos prototipas

Pažymėkite norimus duomenis



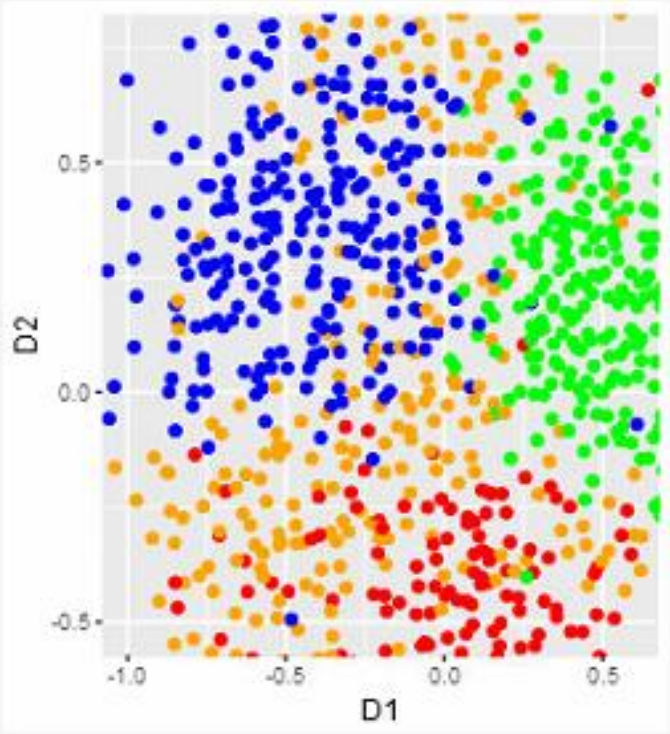
Duomenų peržiūra



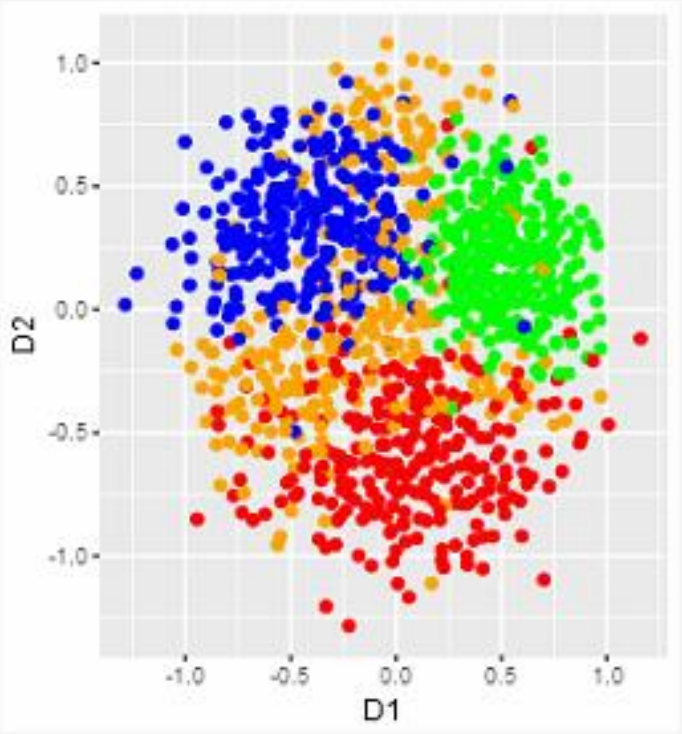
Dimensijų mažinimo metodai

(•) MDS smacof ( ) PCA ( ) ICA ( ) Pricipal Curves ( ) LLE ( ) Isomap

Pažymėkite norimus duomenis



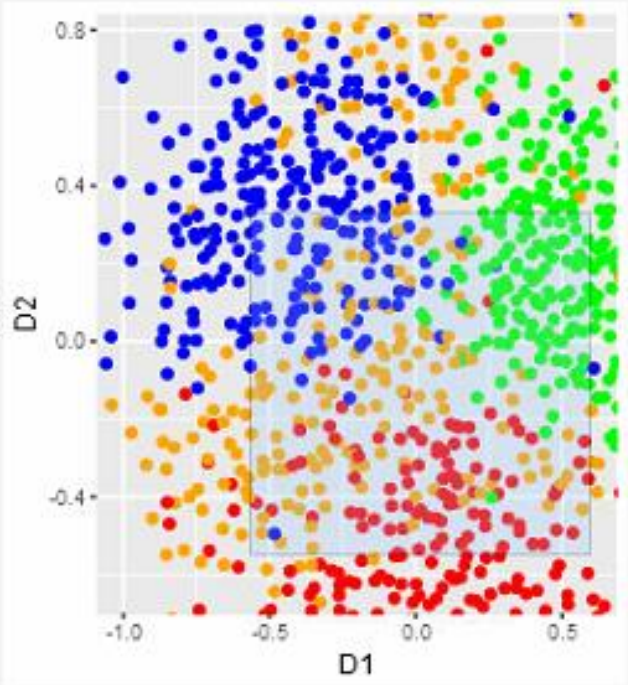
Duomenų peržiūra



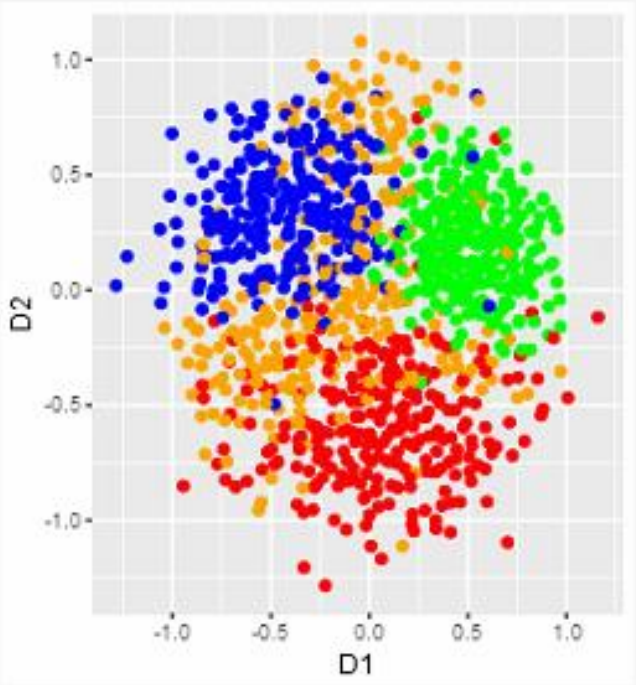
Dimensijų mažinimo metodai

(•) MDS smacof ( ) PCA ( ) ICA ( ) Pricipal Curves ( ) LLE ( ) Isomap

Pažymėkite norimus duomenis



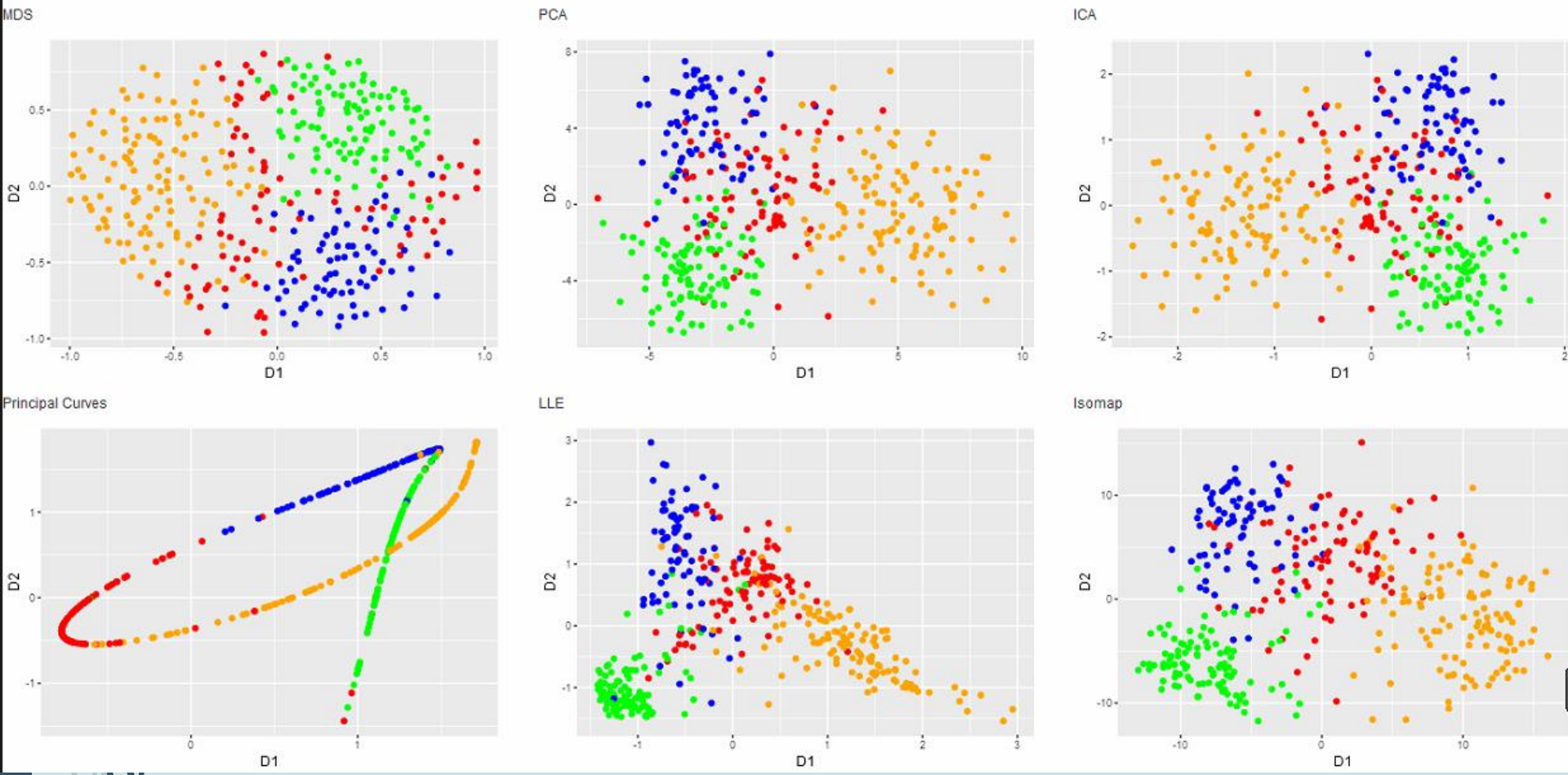
Duomenų peržiūra



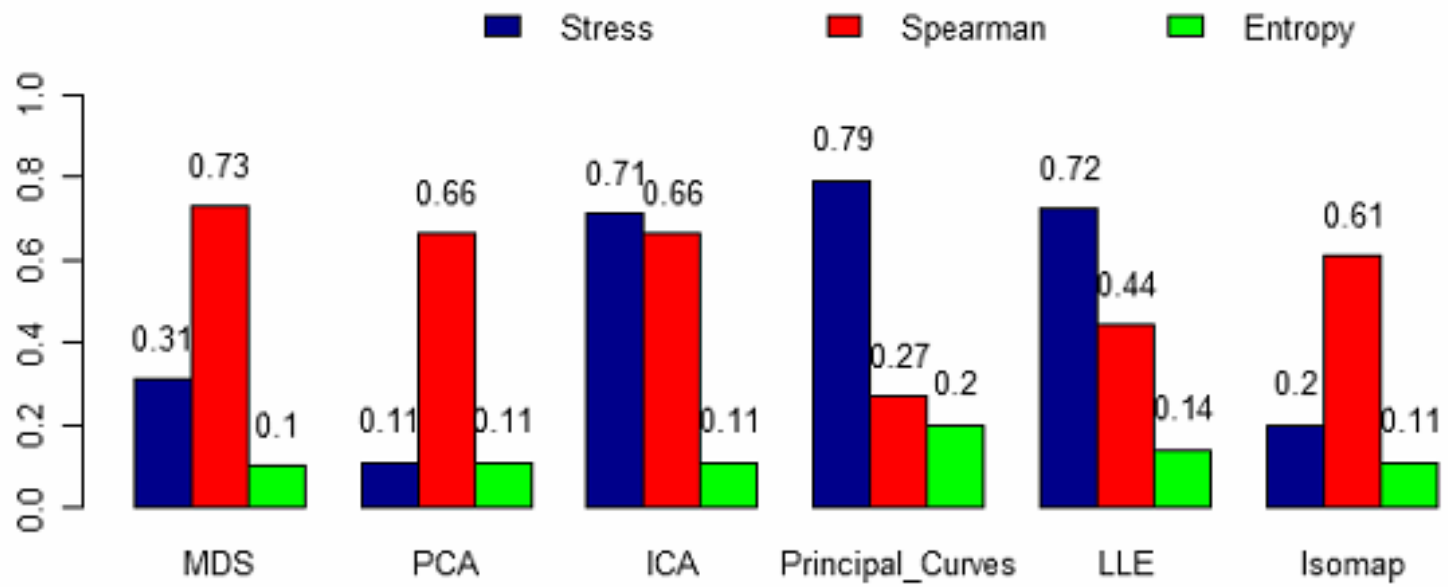
Dimensijų mažinimo metodai

MDS smacof | PCA | ICA | Pricipal Curves | LLE | Isomap

Pradinis vizualizavimas | Vizualizuoti pasirinktu metodu | Vizualizuoti visais metodais


MDS
PCA
ICA
Principal Curves
LLE
Isomap
D1
D2

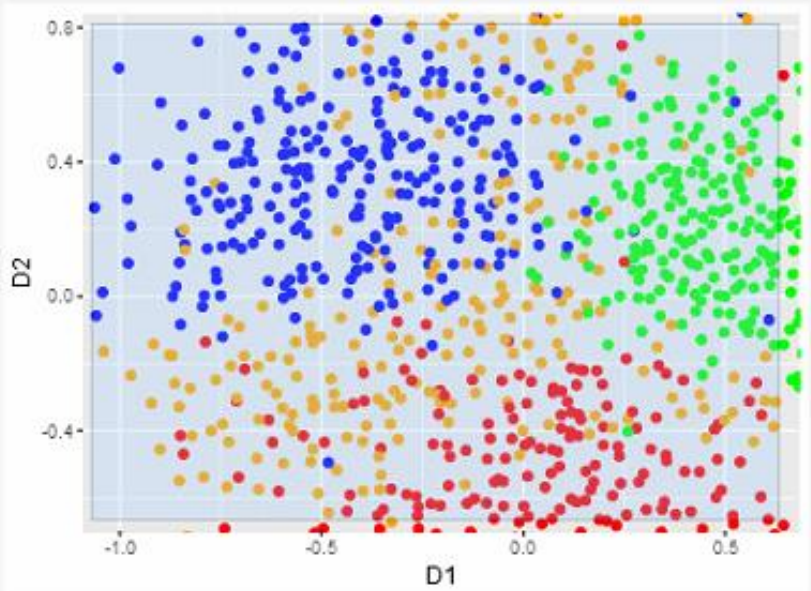
Tiklsumų grafikas



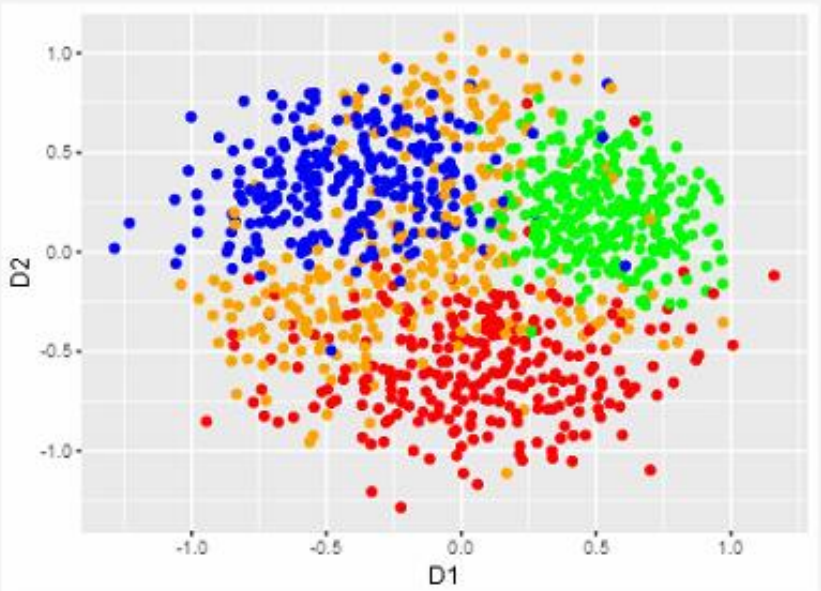
Tikslumai:

	MDS	PCA	ICA	Principal_Curves	LLE	Isomap
Stress	0.31	0.11	0.71	0.79	0.72	0.20
Spearman	0.73	0.66	0.66	0.27	0.44	0.61
Entropy	0.10	0.11	0.11	0.20	0.14	0.11

Pažymėkite norimus duomenis



Duomenų peržiūra

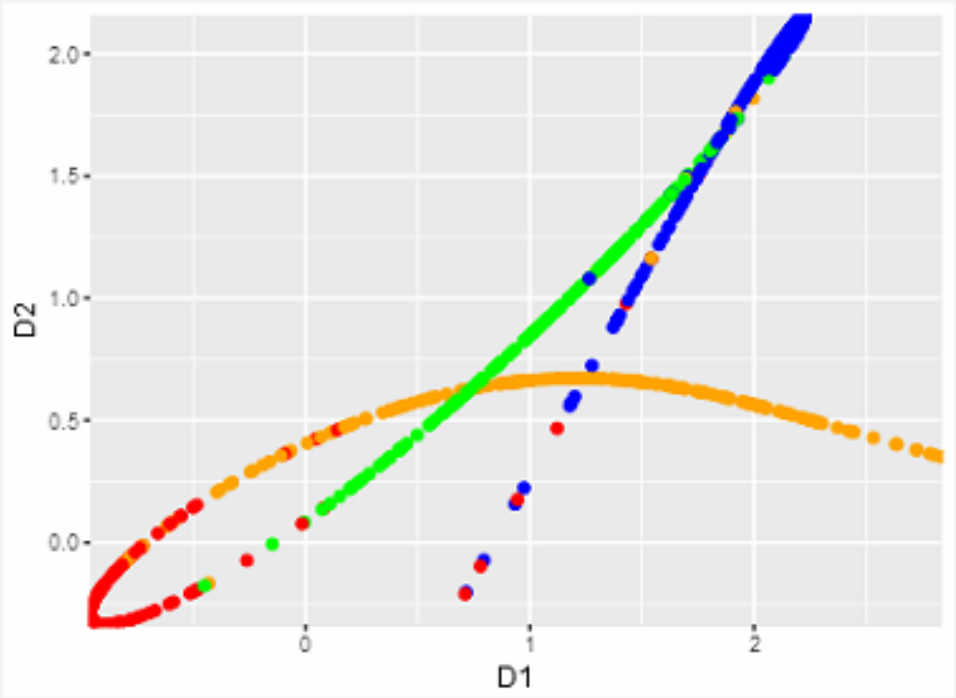


Dimensijų mažinimo metodai

MDS smacof | PCA | ICA | Pricipal Curves | LLE | Isomap

Pradinis vizualizavimas | Vizualizuoti pasirinktu metodu | Vizualizuoti visais metodais

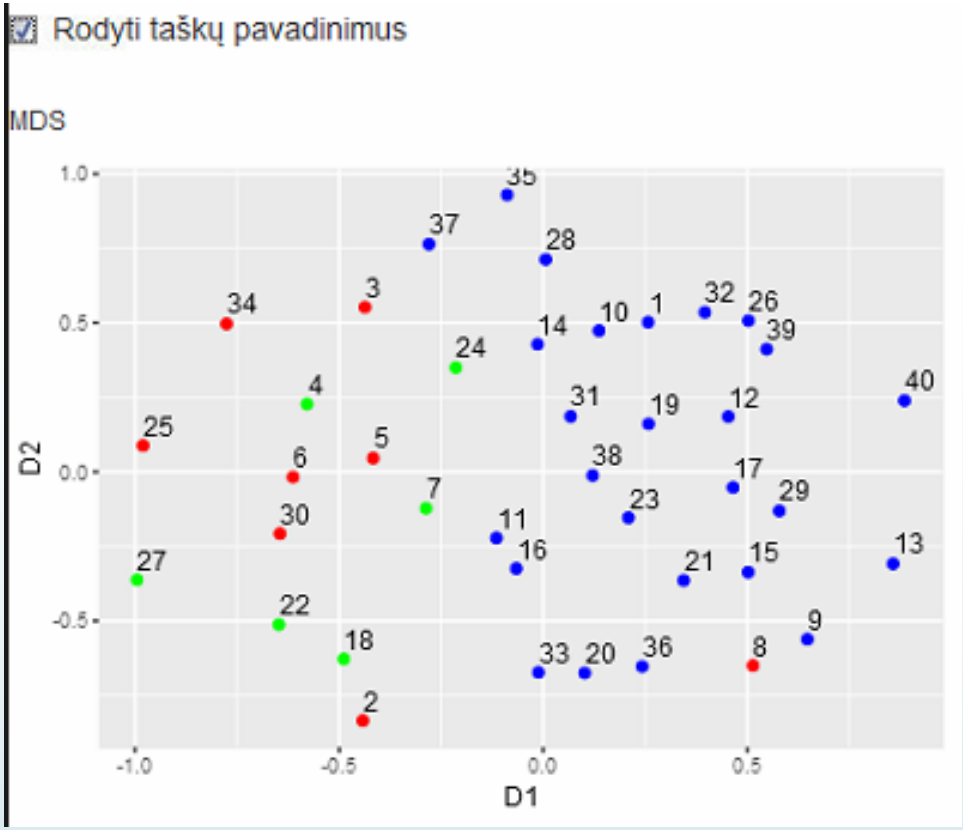
Pažymėkite norimus duomenis



Dimensijų mažinimo metodai

MDS smacof | PCA | ICA | Pricipal Curves | LLE | Iso

Pradinis vizualizavimas | Vizualizuoti pasirinktu metodu | Vizualizu



Atrinktų objektų komponenčių reikšmės

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	0.8001623	1.9781285	-0.9279376	-4.10594955	-1.6292845	-0.4176327	-1.7447228	-1.0907848	0.4486241	-3.5518213
2	0.8495327	2.2089543	0.3802479	-0.40099732	-1.9251724	1.1704299	-1.4473829	1.6577513	1.8363304	-3.6054704
3	-0.5448801	3.6422172	2.2541883	0.40335819	-0.3511837	0.2818392	0.7952896	2.7891934	1.5427610	1.8385038
4	1.1350593	2.9014105	-0.1007830	1.01949532	0.4564027	2.7112218	-0.1850544	-0.9248055	2.9638417	-3.3538990
5	0.5697012	2.1201998	0.7439663	3.93071870	1.9647248	-1.3644770	-3.5469878	-2.2811092	3.1990439	2.6767093
6	1.8539207	-0.2498224	-1.6212033	2.70988180	0.5569820	-3.1440346	1.4339909	-5.4140163	3.9106674	1.7795082
7	0.7895941	-0.3310443	3.2537476	2.14386841	0.3549214	-0.9110015	0.2634310	-6.3510018	2.9725944	-3.1041678
8	0.3000774	-1.0521669	-3.3923346	2.93359327	-5.1334771	-1.2212665	-4.0767909	-1.4074647	0.3208006	4.0056367
9	-0.5089273	5.2972013	0.5018624	0.02110233	0.3080362	2.2532239	0.6645603	2.6728635	1.8880222	2.3778772
10	1.2574222	1.5993354	2.8472655	1.36275339	0.7993932	4.3665894	-4.2919714	-2.0828933	2.2592123	-0.2993374
11	-1.0397251	1.6251207	0.5903140	0.77053642	3.2060194	-2.2817473	-3.1462523	2.6789643	2.1144069	1.6035190
12	-1.0940819	3.8044013	3.6269872	0.14056092	-2.4678865	0.8094776	-0.1405959	1.9271042	1.7983793	2.4755370
13	-0.9308408	-0.8377067	0.0434901	-0.23106326	-0.3010819	-1.5377148	3.3054396	2.5663835	2.2530088	3.5326700
14	-0.4301798	-0.1361530	1.7869641	4.22971721	3.3901783	-2.9616716	0.3457950	0.4450120	2.9724561	3.1960039

2017 – 2018 m. m. darbo planas

- **Mokslinių tyrimų planas:**
  - Gautų rezultatų analizė, apibendrinimas, išvadų parengimas:
    - Gautų rezultatų statistinė analizė;
    - Rezultatų apibendrinimas, esminių rezultatų išskyrimas;
    - Išvadų parengimas.
  - Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų ir kt.) parengimas:
    - Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas;
    - Analitinės disertacijos dalies parengimas;
    - Teorinės disertacijos dalies parengimas;
    - Eksperimentinės disertacijos dalies parengimas;
    - Bendrųjų išvadų formulavimas.

2016 – 2017 m. m. darbo planas

- **Rezultatų pristatymo planas:**
  - Dalyvavimas 7-oje tarptautinėje konferencijoje „Advanced Technology & Sciences (ICAT 2018)“, vyksiančioje 2018 m. rugsėjo mėn. Rygoje, Latvijos Respublikoje.
- **Mokslinių publikacijų planas:**
  - Planuojamas mokslinis straipsnis *Baltic Journal of Modern Computing* žurnale.