Ataskaitinė informatikos krypties doktorantų konferencija
2018-10-24

Didelės apimties duomenų vizuali analizė

Doktorantė: **Jelena Liutvinavičienė**

Darbo vadovė: **Prof. dr. Olga Kurasova**

Doktorantūros pradžia: 2014 m.

Doktorantūros pabaiga: 2018 m.

Temos aktualumas

- Didelių duomenų analizė įgalina atrasti paslėptą informaciją ir panaudoti ją efektyvesnių sprendimų priėmimui.
- Svarbi tokios analizės dalis yra duomenų vizualizavimas, nes būtent jis įgalina pastebėti paslėptus ryšius tarp objektų, ką yra sunku padaryti taikant standartinius analizės metodus.
- Trūksta metodologijos, kuri leistų efektyviausiai panaudoti įvairius metodus greitam ir tiksliam skirtingo dydžio bei pobūdžio duomenų rinkinių atvaizdavimui.

Tyrimo objektas

- Didelės apimties daugiamačiai duomenys.
- Dimensijų mažinimo metodai didelės apimties daugiamačiams duomenims vizualizuoti.

Tyrimo tikslas

- Pasiūlyti integralią didelės apimties duomenų vizualios analizės metodologiją,
  - apimančią skirtingų dimensijų mažinimo metodų taikymą, jų statistinių savybių panaudojimą metodų parinkimui, duomenų paruošimą analizei bei jų klasterizavimą;
  - ir kurią realizuojantis įrankis galėtų veikti paskirstytų skaičiavimų sistemose / debesyje.

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; atlikti metodų greičio ir tikslumo vertinimo tyrimą.
- Pasiūlyti daugiapakopę didelės apimties duomenų vizualizavimo metodologiją; jos veikimo principus (skirtingų dimensijų mažinimo metodų taikymą, statistinių rodiklių panaudojimą, duomenų klasterizavimą) iliustruoti apdorojant skirtingus duomenų rinkinius.
- Sukurti programų sistemos prototipą, kuriame būtų realizuota pasiūlyta didelės apimties duomenų vizualizavimo metodologija.

Gauti rezultatai

- **Darbo mokslinis naujumas:**
  - Atlikta didelių duomenų vizualizacijos metodų ir įrankių lyginamoji analizė.
  - Sukurta didelės apimties duomenų vizualizavimo metodologija.
- **Darbo praktinė reikšmė:**
  - Sukurtas metodologiją realizuojančios sistemos prototipas.
- **Ginamieji teiginiai:**
  - Pasiūlyta metodologija yra tinkama didelės apimties duomenų vizualizavimui pritaikant skirtingus dimensijų mažinimo metodus.
  - Pasiūlyta metodologija leidžia pasirinkti dimensijų mažinimo metodą atsižvelgiant į metodo taikymo greitį ir tikslumą.

Publikacijos periodiniuose leidiniuose 2014–2018 m. m.

- Liutvinavičienė J., Kurasova O. **Multi-level Massive Data Visualization: Methodology and Use Cases.** *Baltic Journal of Modern Computing, Vol 6, No 4 (2018), p. 321-334.*
- Zubova J., Kurasova O., Liutvinavičius M. **Dimensionality reduction methods: the comparison of speed and accuracy.** *Information Technology and Control, Vol 47, No 1 (2018), p. 151-160.* (Žurnalas turi cituojamumo rodiklį *Clarivate Analytics Web of Science* duomenų bazėje, IF 2016: 0.475).
- Zubova J., Kurasova O. (2015). **Didelių duomenų vizualizavimo metodai ir įrankiai.** *Informacijos mokslai* 73: 113–126. Vilnius: Vilniaus universiteto leidykla. ISSN 1392-0561.

Publikacijos konferencijų medžiagoje 2014–2018 m. m.

- 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.
- 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., 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.
- 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.

Kitos publikacijos 2014–2018 m. m.

- 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.*
- 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.

Dalyvavimas konferencijose 2014–2018 m. m.:

- 2017-09-21 – 2017-09-22 Lietuvos kompiuterininkų sąjungos organizuotas renginys ***„Kompiuterininkų dienos – 2017“***, vykęs XVIII kompiuterininkų konferencijoje, Kaunas, pranešimas ***„Daugiamatiškumo mažinimo metodai: greičio ir tikslumo palyginimas“***.
- 2016-12-01 – 2016-12-03 7-oje mokslinė konferencija ***„Duomenų analizės metodai programų sistemoms“***, Druskininkai, stendinis pranešimas ***„Multi-level Method for Big Data Visualization“***.
- 2016-10-13 – 2016-10-15 tarptautinė konferencija ***„22nd International Conference on Information and Software Technologies“*** (ICIST 2016), Kaunas, pranešimas ***„Parallel computing for dimensionality reduction“***.
- 2016-08-08 – 2016-08-12 5th GESIS Summer School in Survey Methodology, Kelnas, Vokietija, išklausytas kursas ***„Introduction to Data Analysis Using R“***.

Dalyvavimas konferencijose 2014–2018 m. m.:

- 2016-04-14 – 2016-04-16 4-oje tarptautinė konferencija ***„Computational Economics and Finance“***, Paryžius, stendinis pranešimas ***„Financial crisis prediction: behavioural finance approach for stock market forecasting“***.
- 2015-12-03 – 2015-12-05 7-oji mokslinė konferencija ***„Duomenų analizės metodai programų sistemoms“***, Druskininkai, stendiniai pranešimai ***„What is Big Data“, „Visual Analytics for Big Data“***.
- 2015-09-17 – 2015-09-19 Lietuvos kompiuterininkų sąjungos organizuotaa renginys „Kompiuterininkų dienos 2015“, vykęs XVII kompiuterininkų konferencijoje, Panevėžys, pranešimas ***„Didelių duomenų vizualizavimo metodai ir įrankiai“***.
- 2014-12-04 – 2014-12-06 6-oji mokslinė konferencija ***„Duomenų analizės metodai programų sistemoms“***, Druskininkai, stendinis pranešimas ***„Challenges of Big Data Visualization“***.

Atlikti tyrimai

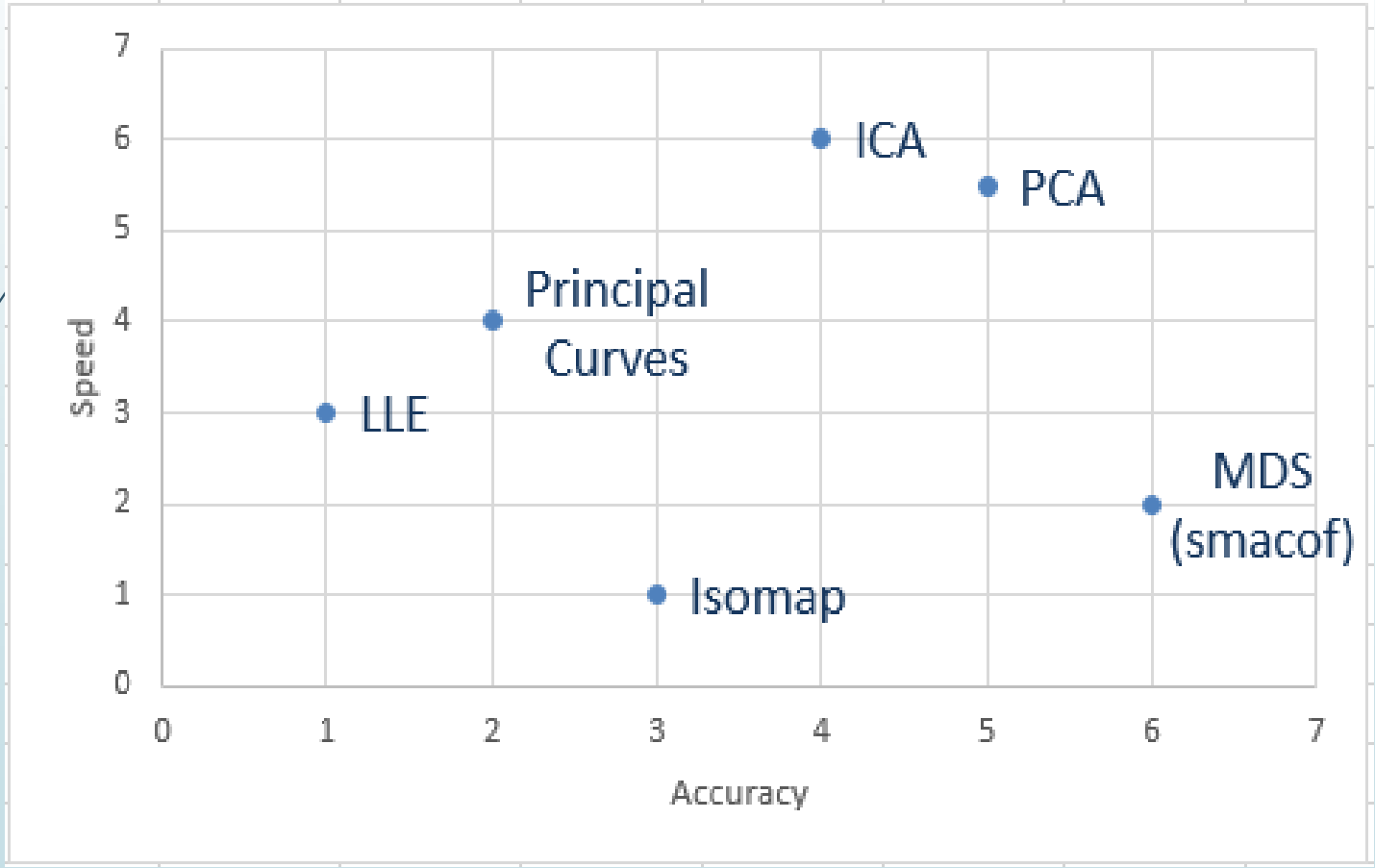
Dimensijų mažinimo metodų greičio ir tikslumo įvertinimas

- Buvo atliktas 6 dimensijų mažinimo metodų algoritmų vykdymo greičio ir tikslumo tyrimas.
- Tikslumas vertintas pagal 3 kriterijus: Stress reikšmę, Spearman‘o koeficientą ir Shannon entropiją.
- Dimensijų mažinimas atliktas su skirtingo tipo duomenimis: atsitiktinai sugeneruotais neklasterizuotais duomenimis, atsitiktinai sugeneruotais klasterizuotais duomenimis ir realiais finansiniais duomenimis.

Dimensijų mažinimo metodų greičio ir tikslumo įvertinimas

- Atsitiktinai sugeneruotų duomenų atvejais (klasterizuotiems ir neklasterizuotiems) buvo patvirtinta keletas taisyklių:
  - Didėjant objektų skaičiui, vykdymo laikas ilgėja. Tačiau pradinis dimensijų kiekis neturi reikšmingos įtakos greičiui.
  - Tikslumui galioja priešingos taisyklės. Didesnis objektų skaičius neturi įtakos tikslumui, tačiau didėjant pradinių dimensijų kiekiui, tikslumas mažėja.
- Realių duomenų atveju pastebėta, jog pradinis dimensijų kiekis gali turėti nedidelės įtakos vykdymo laikui. Taip pat šių duomenų nepavyko apdoroti LLE metodu ir gauti Stress reikšmės ICA metodui. Tai rodo, jog realių duomenų apdorojimas yra labiau komplikuotas. Rezultatai parodė, kad didesnis objektų kiekis šiuo atveju lemia didesnį tikslumą.

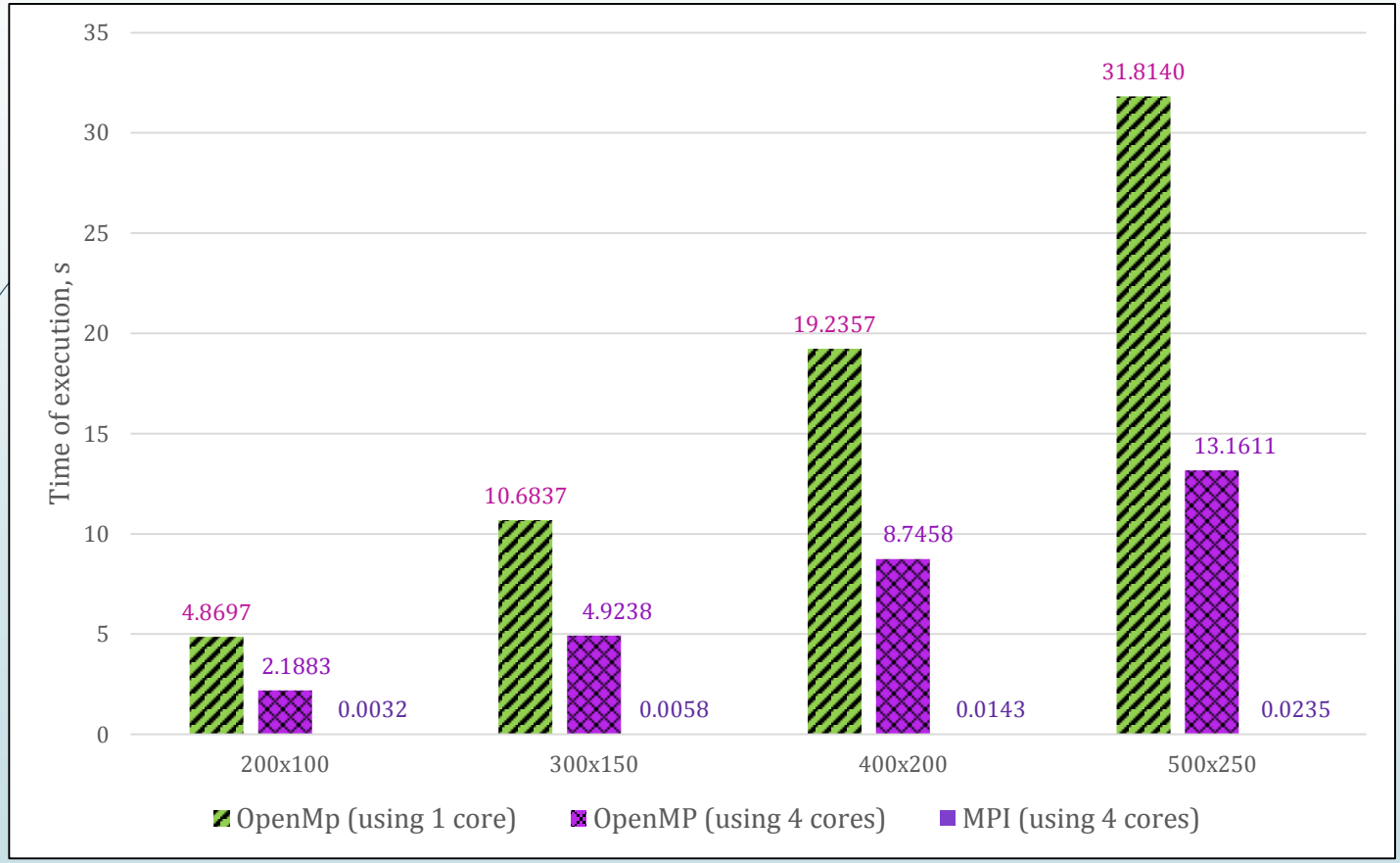
Metodų palyginimas atsitiktinai sugeneruotų neklasterizuotų duomenų atveju



Infrastruktūros įtaka duomenų apdorojimo spartai

- Tyrime analizuota, kaip lygiagrečiųjų skaičiavimų taikymas gali paspartinti duomenų dimensijų mažinimo bei vizualizavimo užduotis.
- Atlikti bandymai parodė, kad vykdant lygiagretiems procesams pritaikytą OpenMP kodą personaliniame kompiuteryje užduotys yra įvykdomos keletą kartų greičiau nei naudojant nuoseklų kodą.
- Dimensijų mažinimą vykdant kompiuterių klasteryje panaudojant MPI technologiją programų vykdymo laikas buvo šimtus kartų trumpesnis, nei tai darant personaliniame kompiuteryje su OpenMP kodu.
- Dimensijų mažinimas panaudojant lygiagrečius skaičiavimus buvo išbandytas realių finansinių duomenų vizualizavimui. Tas užduotis, kurių buvo nebeįmanoma atlikti su asmeniniu kompiuteriu, panaudojant MPI technologiją, pavyko įgyvendinti greičiau nei per 0,4 sekundės.

Infrastruktūros įtaka duomenų apdorojimo spartai

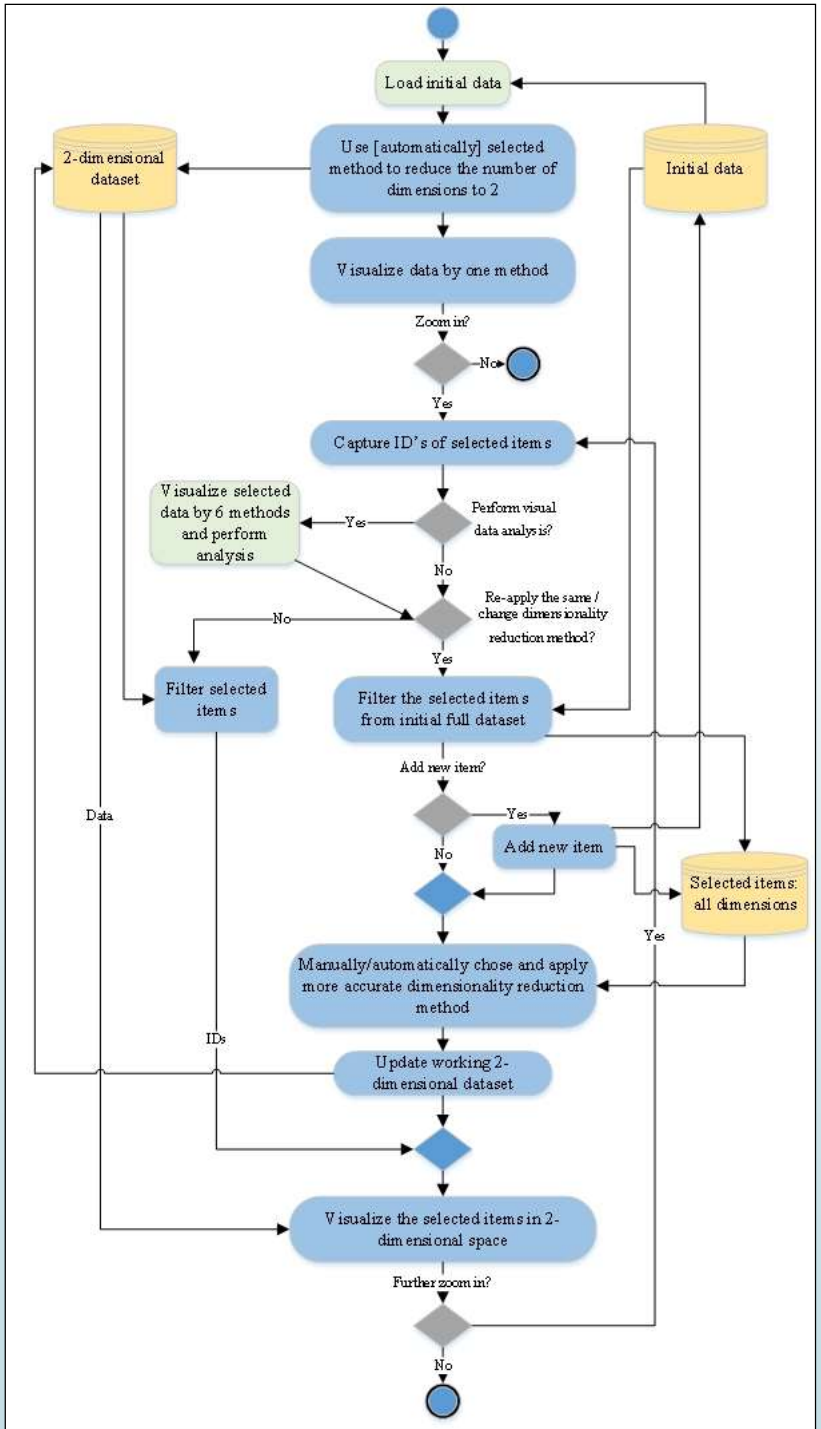


Dimensijų mažinimas atliktas panaudojus *Random projection* metodą

Siūloma metodologija

- Duomenų vizualizavimas yra paremtas dimensijų mažinimo metodais.
- Vizualizavimo procesas suskaidomas į atskirus žingsnius:
  - Kiekviename žingsnyje tam tikras dimensijų mažinimo metodas gali būti pritaikytas atsižvelgiant į duomenų tipą ir kiekį.
  - Metodai parenkami pagal jų algoritmų vykdymo greitį ir tikslumą.
  - Kai duomenys apdorojami ir vizualizuojami, yra galimybė peržiūrėti visų klasterių parametrų statistinius duomenis.
  - Tolesnę analizę/vizualizavimą galima atlikti tik su pasirinktais duomenų elementais (grafike pažymint norimą plotą).

Metodologijos algoritmo schema



Metodologijoje naudojami dimensijų mažinimo metodai

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

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.

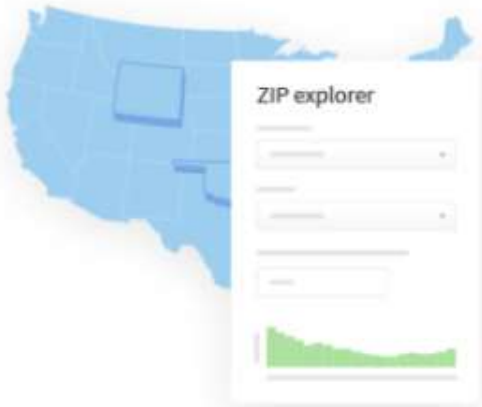
Siūlomos metodologijos prototipas

Duomenų vizualizavimo įrankių apžvalga

RStudio®

RStudio

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



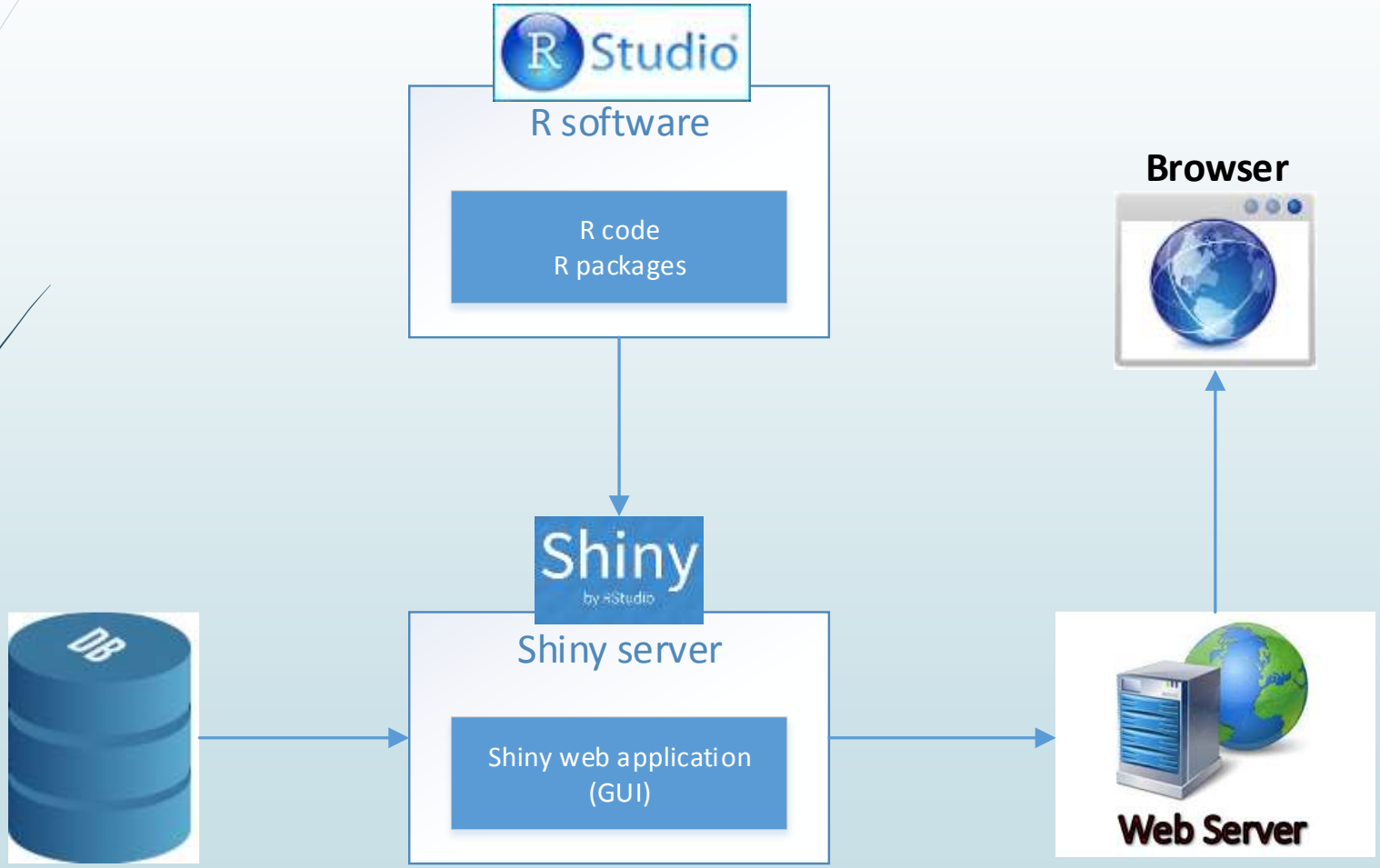
Shiny

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

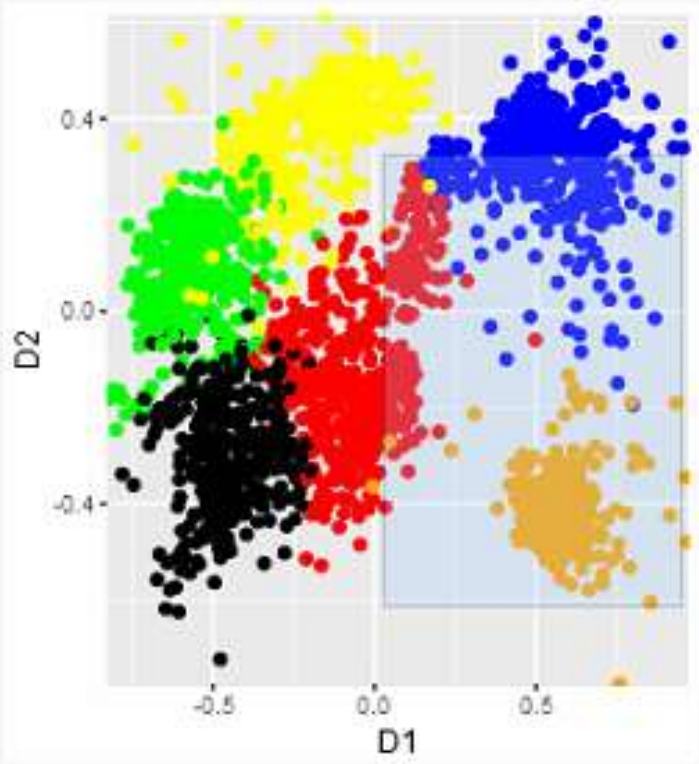
R Packages

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

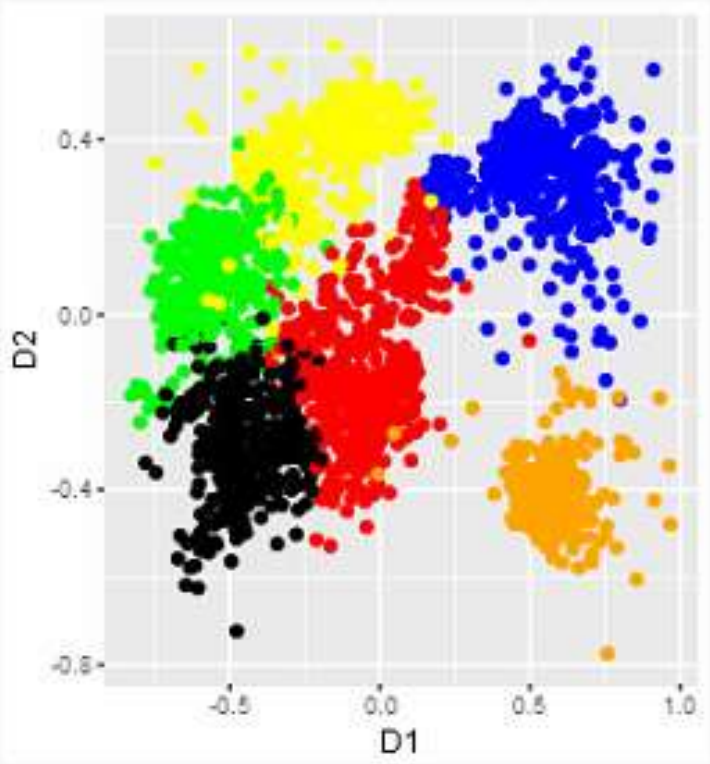
Įrankio architektūra



Pažymėkite norimus duomenis



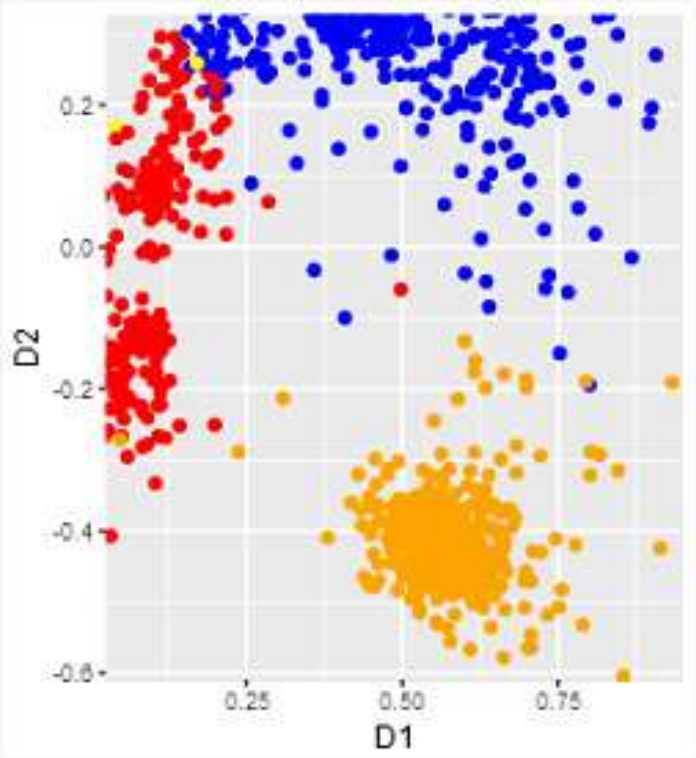
Duomenų peržiūra



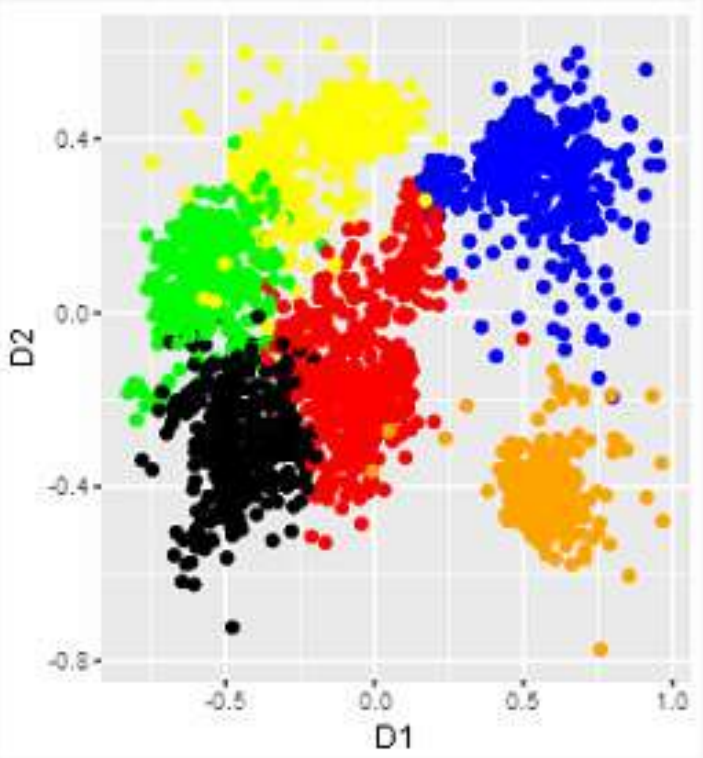
Dimensijų mažinimo metodai

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

Pažymėkite norimus duomenis



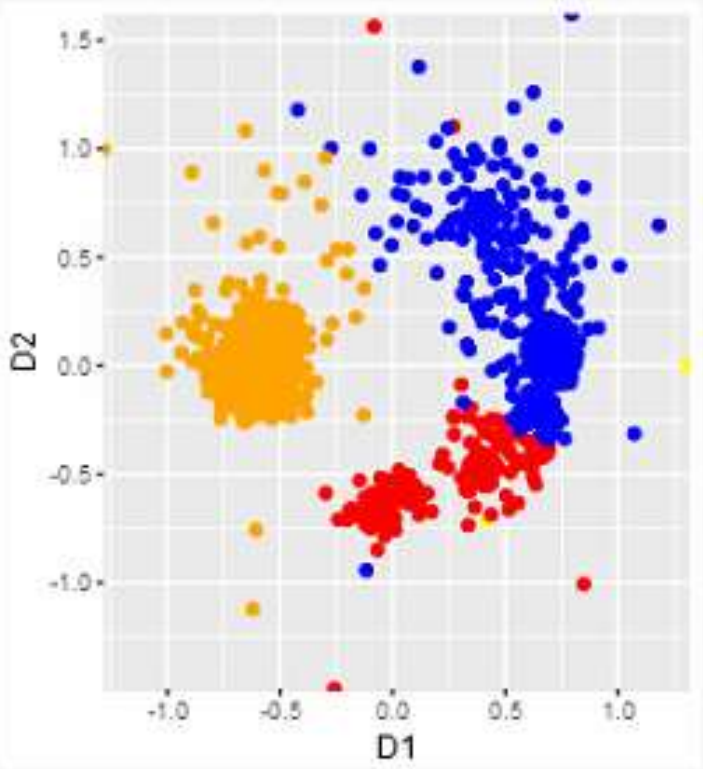
Duomenų peržiūra



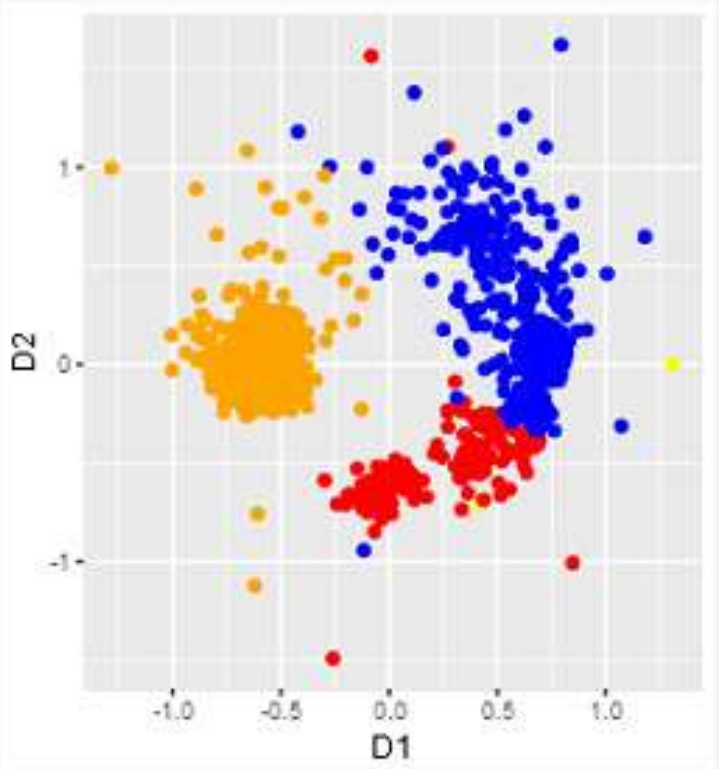
Dimensijų mažinimo metodai

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

Pažymėkite norimus duomenis



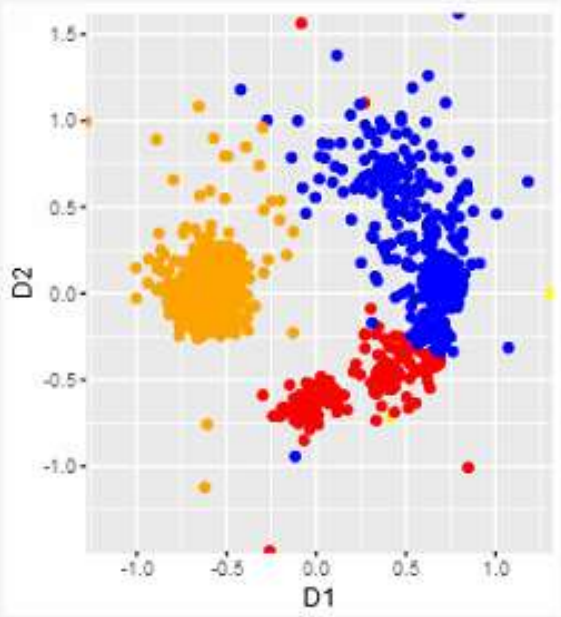
Duomenų peržiūra



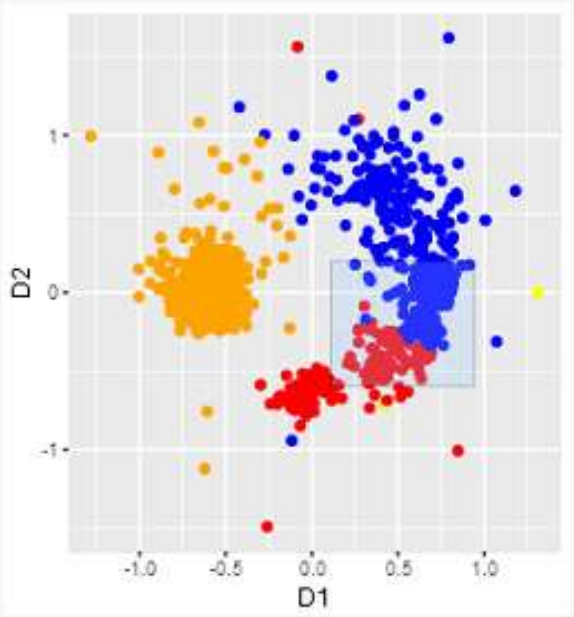
Dimensijų mažinimo metodai

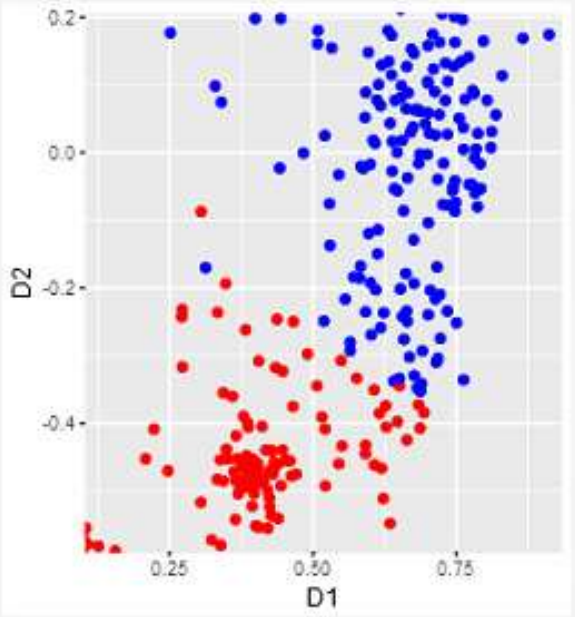
(•) MDS smacof ( ) PCA ( ) ICA ( ) Pricipal Curves ( ) LLE ( ) Isomap

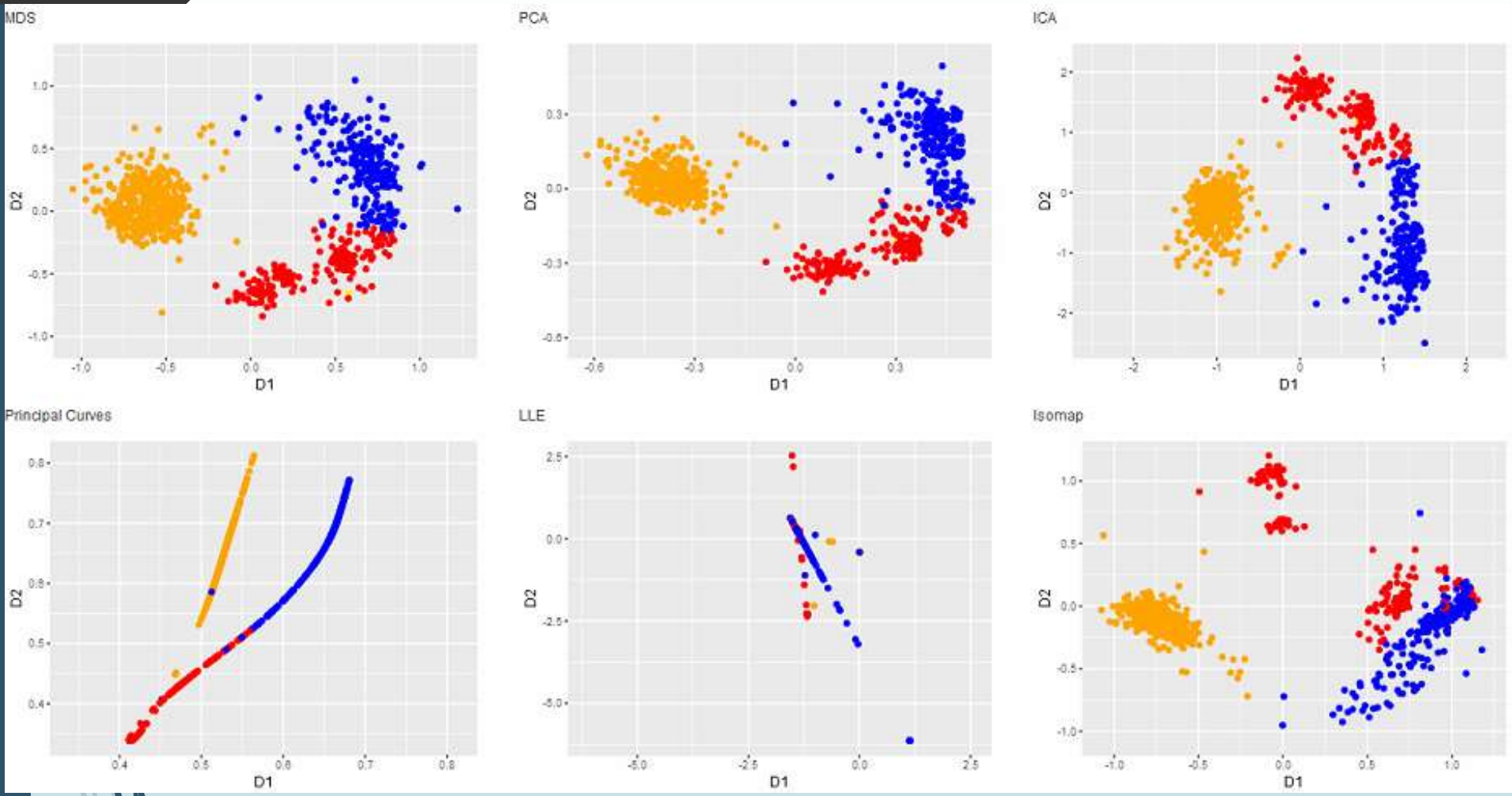
Pažymėkite norimus duomenis



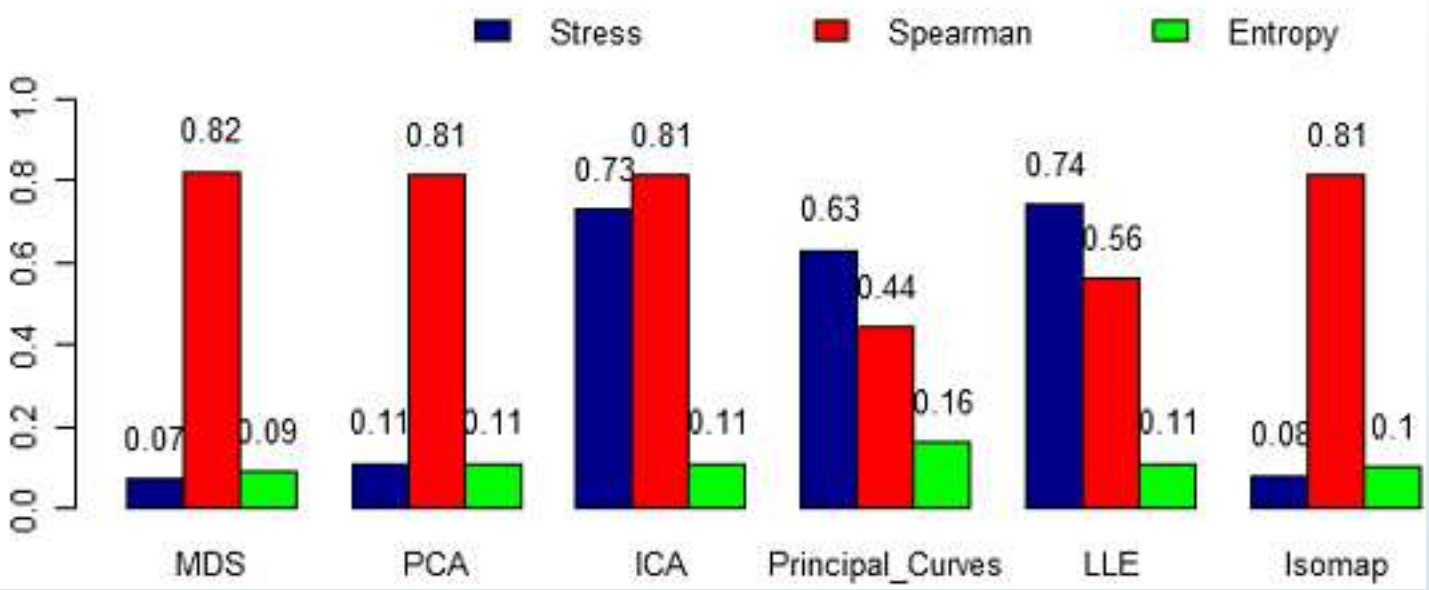
Duomenų peržiūra






MDS
PCA
ICA
Principal Curves
LLE
Isomap
D1
D2

Tikslumų grafikas

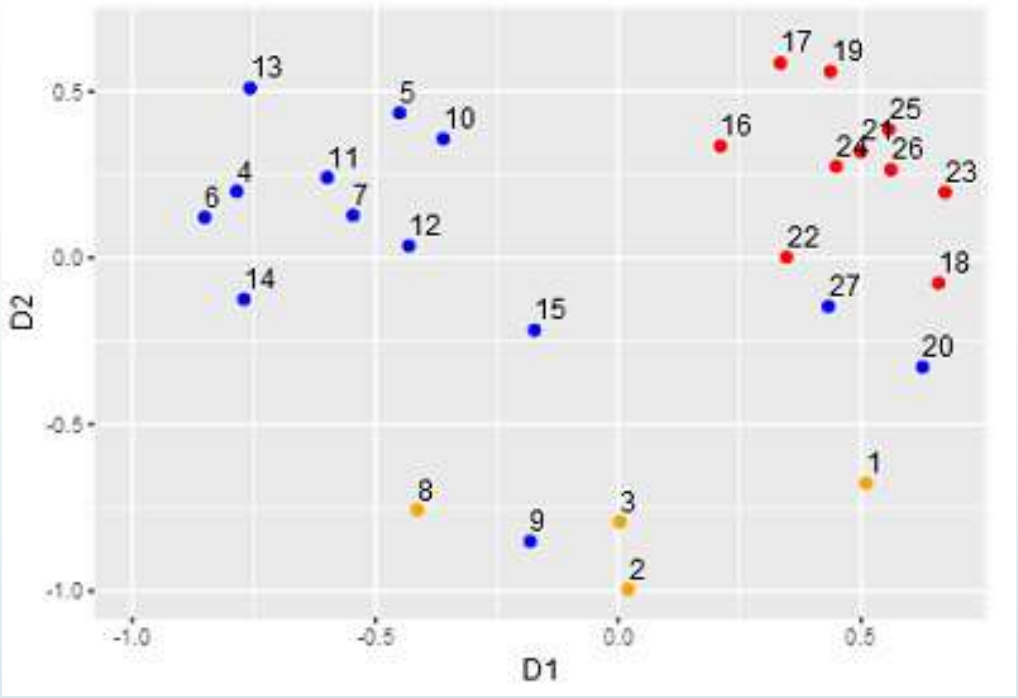


Tikslumai:

	MDS	PCA	ICA	Principal_Curves	LLE	Isomap
Stress	0.07	0.11	0.73	0.63	0.74	0.08
Spearman	0.82	0.81	0.81	0.44	0.56	0.81
Entropy	0.09	0.11	0.11	0.16	0.11	0.10

☑ Rodyti taškų pavadinimus

MDS



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

Rezultatai ir išvados

- Pasiūlyta metodologija suteikia daugiau galimybių vizualizuojant didelius duomenis. Ji leidžia duomenis analizuoti įvairiais pjūviais, duomenis analizuoti pažingsniui, kiekviename etape pasirenkant dominančią duomenų aibę ir jai pritaikant geriausiai tinkantį dimensijų mažinimo metodą, atsižvelgiant į dimensijų mažinimo greitį ir tikslumą konkrečiu atveju.
  - Metodų pasirinkimą palengvina pateikiami grafiniai vizualizavimo pavyzdžiai bei greičio ir tikslumo rodikliai.
- Sukurtas įrankis, kuris realizuoja pasiūlytą metodologiją. Įrankio architektūra įgalina duomenų apdorojimą atlikti panaudojant debesų kompiuterijos išteklius.