Vilniaus Universitetas

Matematikos ir Informatikos fakultetas

Daugiamačių duomenų vizualizavimas

2 Užduotis
- Papildymas -

Tiesioginio vizualizavimo metodai

Laimonas Beniušis
Informatika 1gr.

Turinys

1 Visualizavimo metodai

- *Attribute Graph*
- *Scatter Plot*
- *Andrew Curves*
- *FreeViz*
- *RadViz*
- *Parallel Coordinates*

2 Atributai

Duomenų aibės atributai yra ženkliai sumažinti ik ~ (770 ir 150) (buvo 5172). Ne visi atributai yra tinkami vizualizavimui, todėl palikti tik šie:

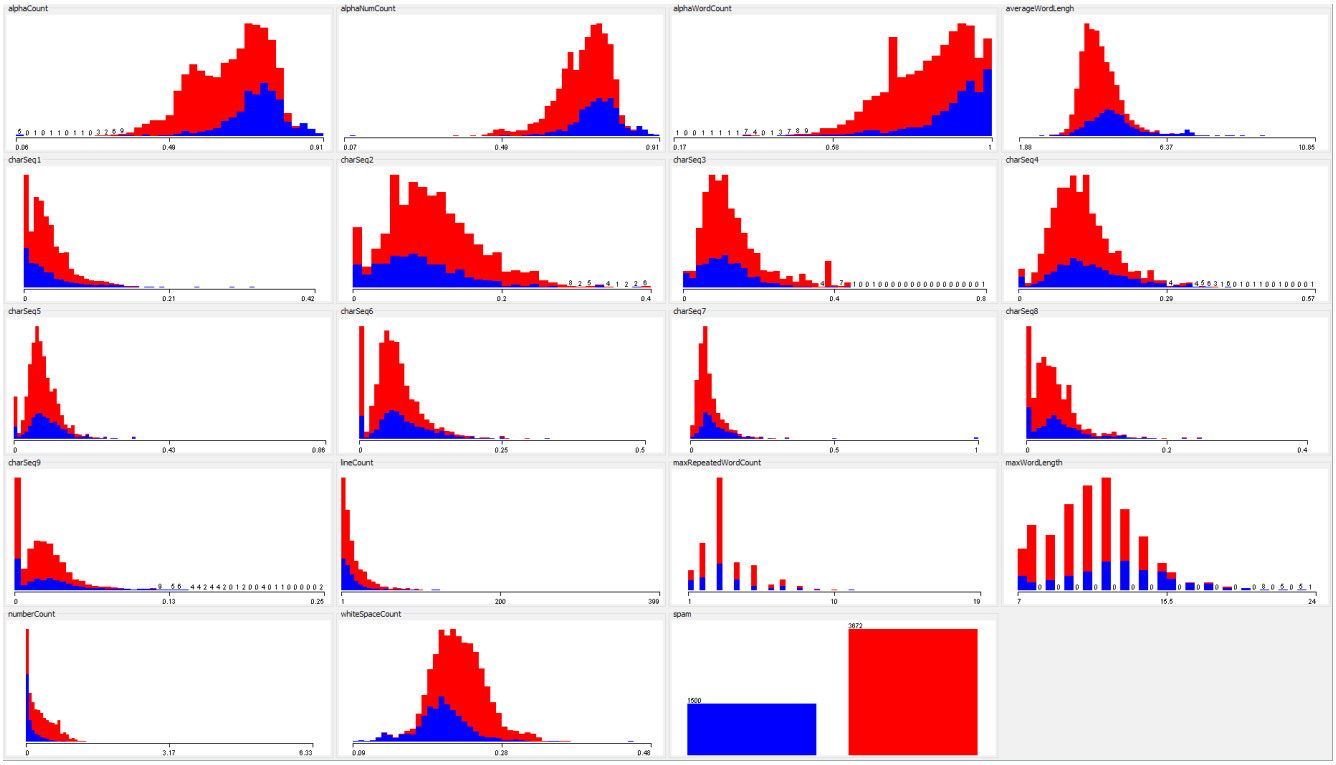
alphaCount	Raidžių kiekis / simbolių kiekis
alphaNumCount	Raidžių ar skaičių kiekis/ simbolių kiekis
alphaWordCount	Raidinių žodžių kiekis / žodžių kiekis
charSeq1	Raidinių žodžių kiekis, kurių ilgis 1 / žodžių kiekis
charSeq2	Raidinių žodžių kiekis, kurių ilgis 2 / žodžių kiekis
charSeq3	Raidinių žodžių kiekis, kurių ilgis 3 / žodžių kiekis
charSeq4	Raidinių žodžių kiekis, kurių ilgis 4 / žodžių kiekis
charSeq5	Raidinių žodžių kiekis, kurių ilgis 5 / žodžių kiekis
charSeq6	Raidinių žodžių kiekis, kurių ilgis 6 / žodžių kiekis
charSeq7	Raidinių žodžių kiekis, kurių ilgis 7 / žodžių kiekis
charSeq8	Raidinių žodžių kiekis, kurių ilgis 8 / žodžių kiekis
charSeq9	Raidinių žodžių kiekis, kurių ilgis 9 / žodžių kiekis
numberCount	Skaičių (žodžių iš skaičių) kiekis / žodžių kiekis
whiteSpaceCount	Tarpinių simbolių (*whitespace*) kiekis / simbolių kiekis
spam	{SPAM,NOTSPAM} E.laiško kategorija

Su WEKA įrankiais (Taškinių grafikų matrica ir atributų grafikai) buvo naudojami visi atributai, o toliau buvo palikti atributai, kuria fiksuoja tam tikro ilgio žodžių kiekį ir tarpinių simbolių kiekį (išimti alphaCount, alphaNumCount, alphaWordCount, numberCount), norint išvadas remiantis natūralios kalbos žodių ilgio dažnumais.

3 Duomenų vizualizavimas

3.1 Attribute graph (atributų grafikai)

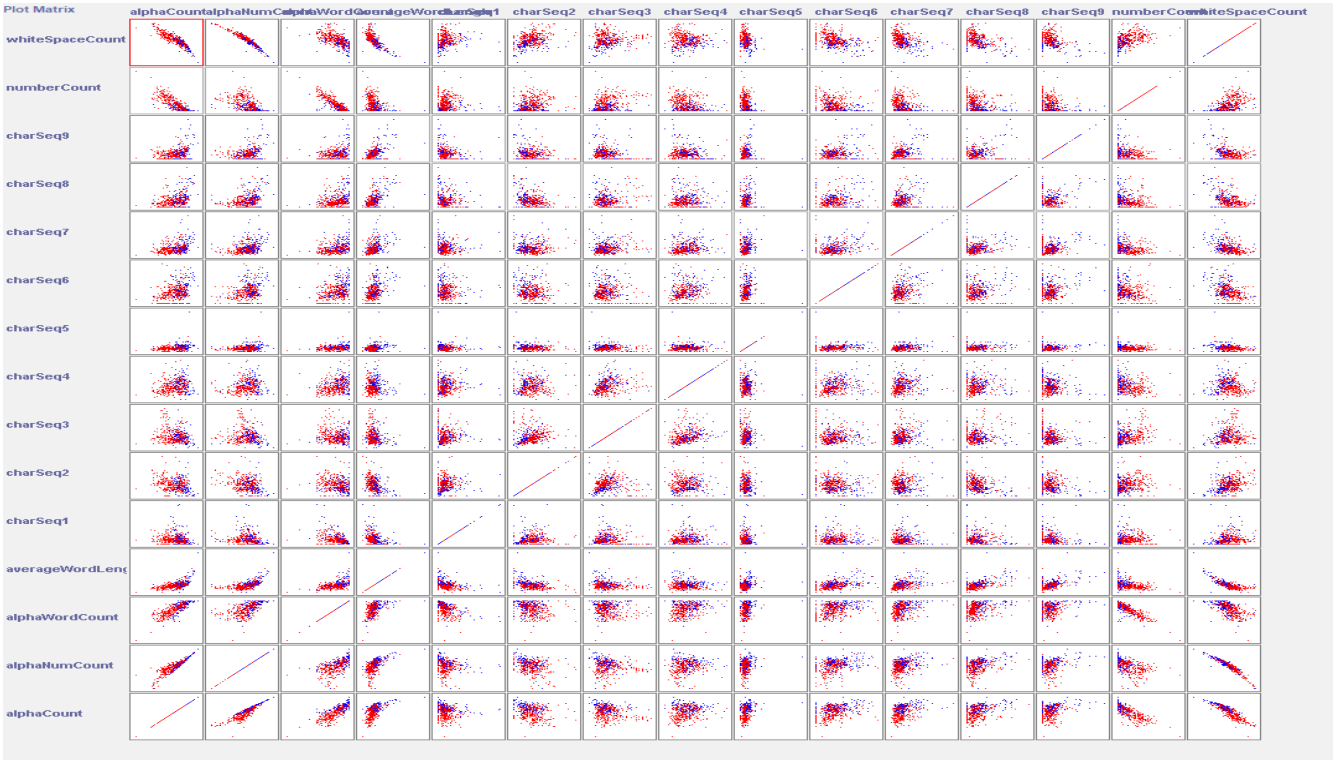
Pats paprasčiausias duomenų vizualizavimo metodas yra juos tiesiog pavaizduoti stulpelinėmis diagramomis. Tada galima pamatyti kaip yra pasiskirsčiusios atributų reikšmės atitinkamoje klasėje. Šiuo atveju yra paliekami visi atributai, nes duomenys yra lyginami neišeinant iš atributo konteksto. Matosi, kad skirtingų duomenų nėra vienodas kiekis.



1. Pav: Weka duomenų attributų grafikai (5172 įrašai)

3.2 Scatter Plot (taškinių grafikų matrica)

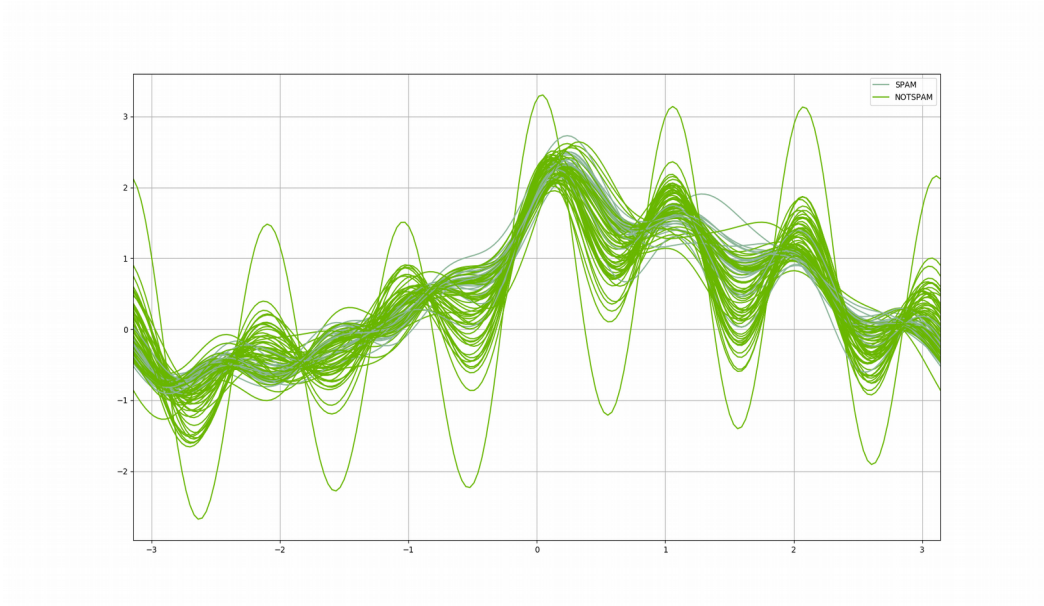
Taškinių grafikų matricoje galima pamatyt, kaip kos duomuo yra susijęs tiek su kitu, tiek kaip pačio atributo duomenys yra pasiskirstę. Pavyzdžiui, jeigu laiške yra daug žodžių (alphaWordCount), akivaizdu, kad bus ir daug raidžių. Taip pat galima pamatyti, kad tarpų kiekis (whiteSpaceCount) yra atvirkščiai proporcingas tik raidžių (alphaCount) ir skaičių su raidėmis (alphaNumCount). Be to, matosi kad žodžių iš 6 raidžių iš viso yra nedaug (santykinai su kitokio ilgio žodžiais).



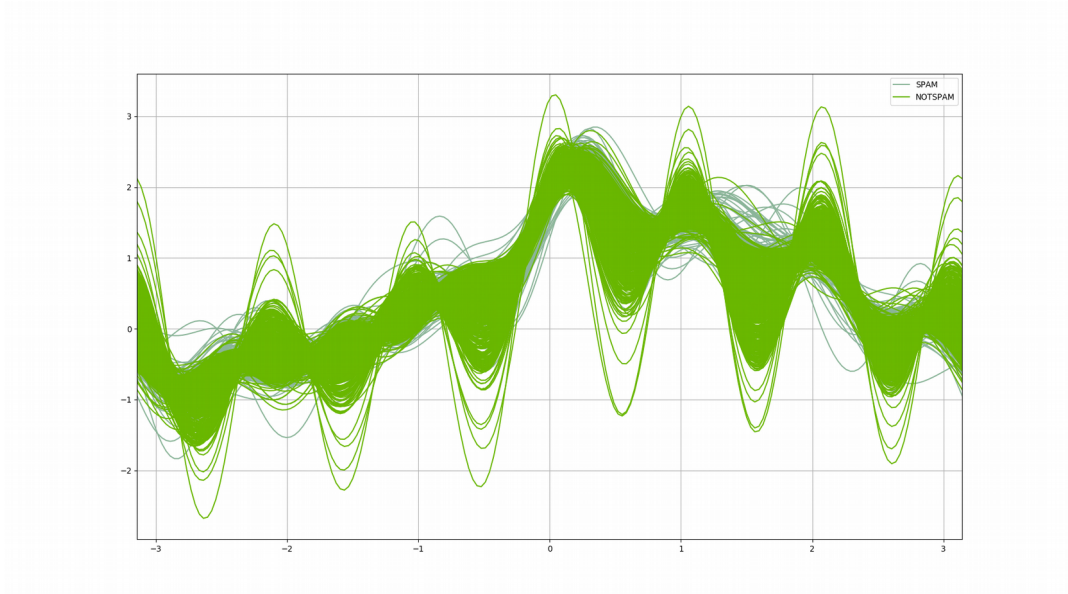
2. Pav: Weka duomenų Scatter Plot matrica

3.3 Andrew Curves (Andrew kreivės)

Andrew kreivės padeda vizualizuoti, kada duomenų kiekis yra didelis, tačiau juos suvokti ir interpretuoti yra gana sunku. Šiuo atveju galime pamatyti, kad abi kategorijos yra panašiai pasiskirsčiusios, tačiau labai akivaizdžiai matosi (žr. 3 Pav. ir 4 Pav.), kad yra įrašų (NOTSPAM kategorijos) kurie žymiai atitolę nuo visomus.



3. Pav: Pandas duomenų Andrew kreivės, 150 įrašų

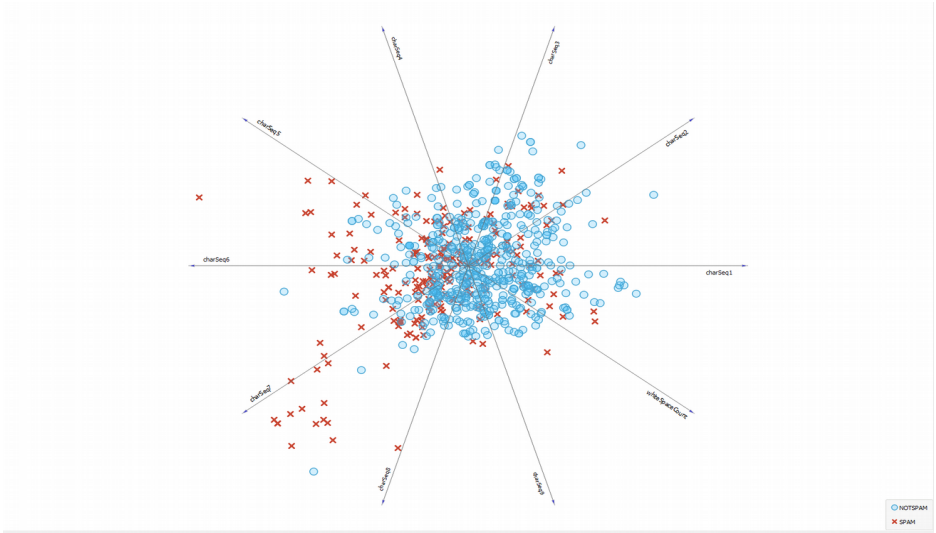


4. Pav: Pandas duomenų Andrew kreivės, 770 įrašų

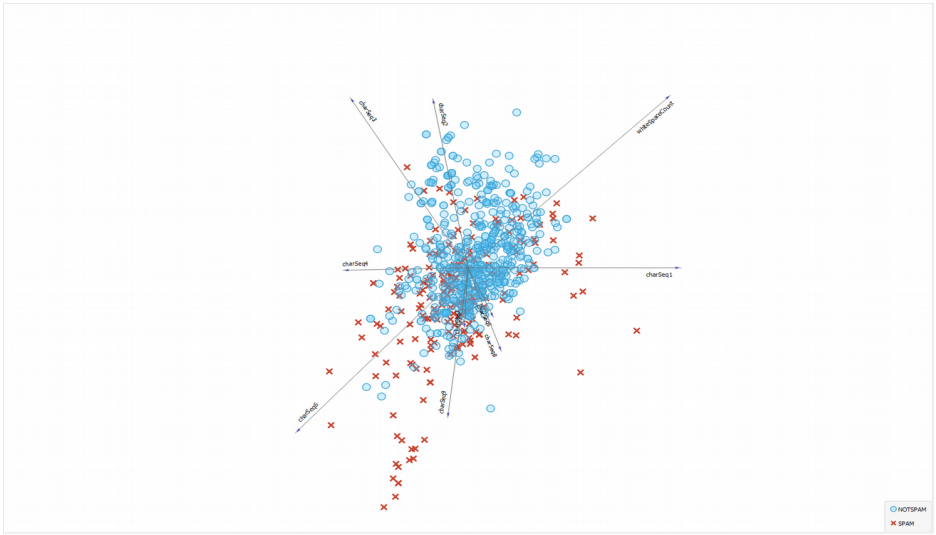
3.4 Spindulinis vizualizavimas

3.4.1 FreeViz

Spinduliniame vizualizavimo metode galima grafiškai pastebėti prie kokių atributų telkiasi klasės atributas. Iš grafikų (žr 5 Pav. ir 6 Pav.) galima pamatyti, kad dauguma laiškų, kurių ilgų žodžių (7-9 simbolių ilgis) santykis labiau pasitaiko prie SPAM kategorijos. Tai yra gana natūrali išvada, nes natūralioje kalboje, ilgi žodžiai yra reti. Pirmame grafike (5. Pav) atributai yra vienodai nutolę vienas nuo kito, o antrame (6 Pav.) atributai yra nutolę atsižvelgiant į duomenis ir įrašų kiekį ir atributo dažnumą juose.



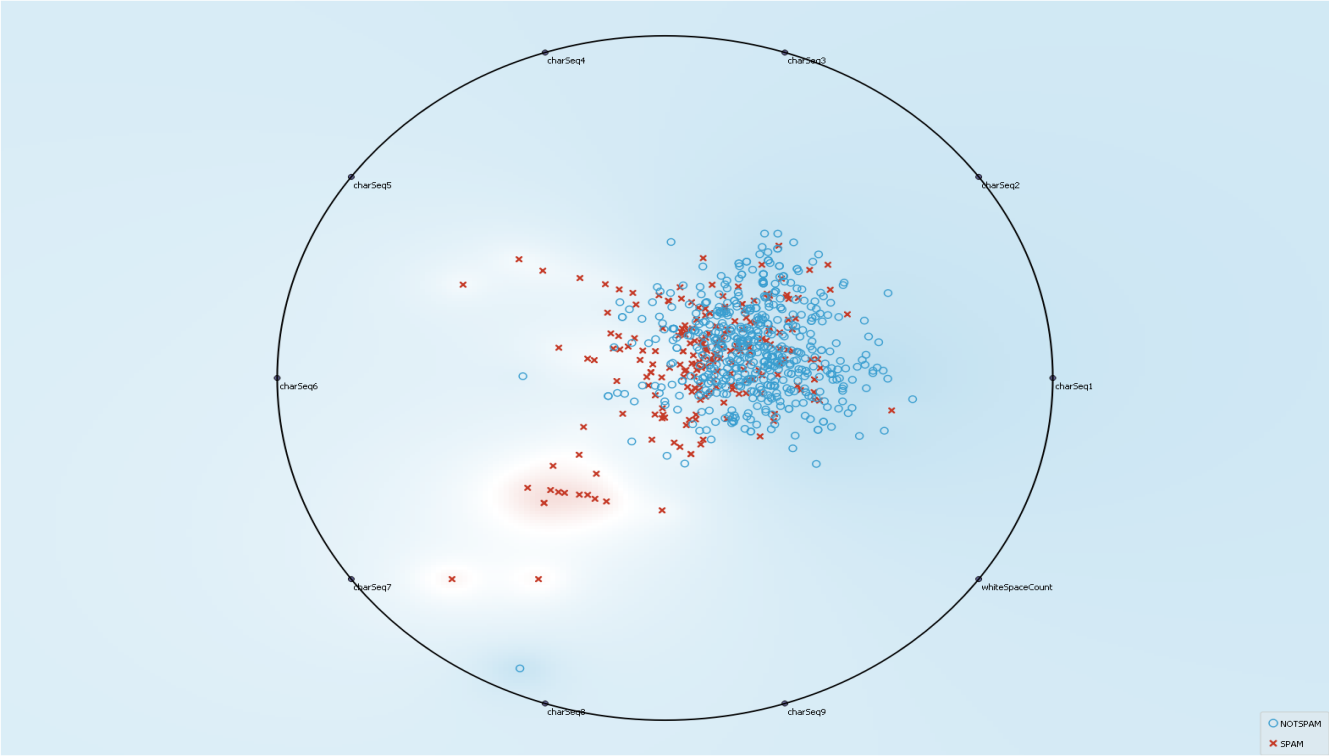
5. Pav: Orange duomenų FreeWiz vizualizacija (neoptimizuota), 770 įrašų



6. Pav: Orange duomenų FreeWiz vizualizacija (optimizuota), 770 įrašų

3.4.2 RadViz

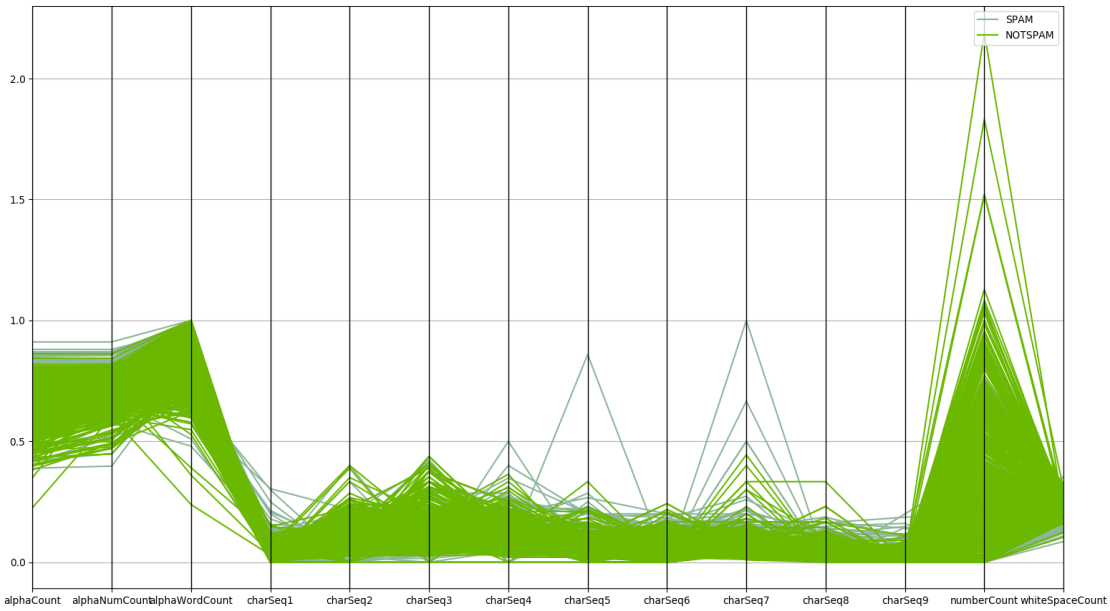
Čia analogiškai kaip ir ankstesniuose spindulinio vaizdavimo grafikuose matome tendenciją, kad didelis ilgų žodžių santykis tikriausiai reiškia, kad el. laiškas yra brukalas.



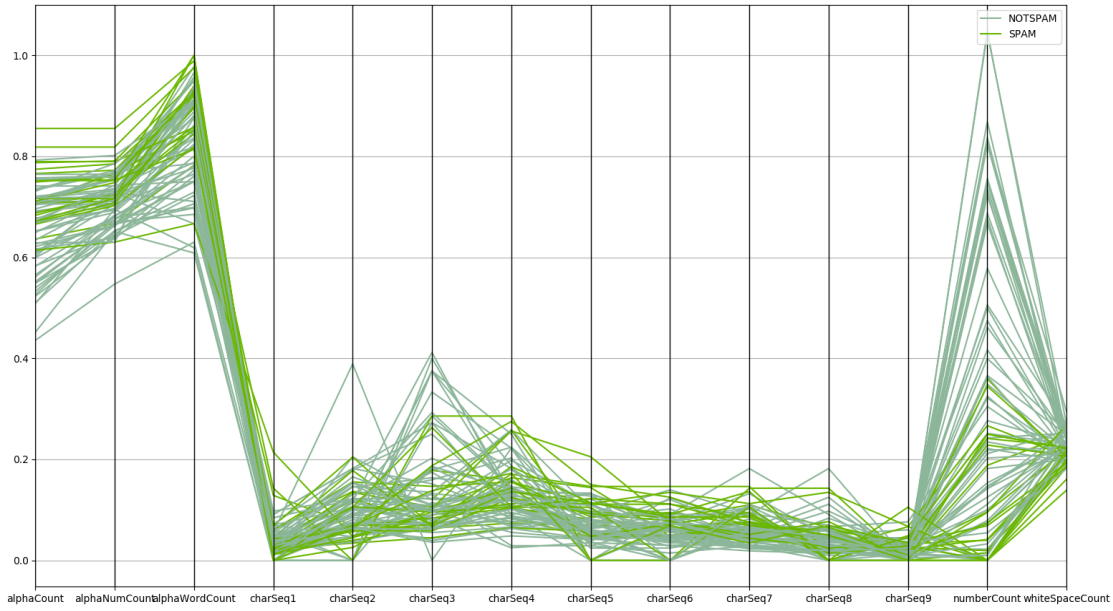
7. Pav: Orange duomenų RadViz vizualizacija , 770 įrašų

3.5 Parallel Coordinates (lygiagrečių kooridinačių vizualizacija)

Lygiagrečių koordinačių vaizdavimo metodas, analogiškai kaip ir Andrew kreivių metodas, nedavė daug naudos, nes duomenys yra per ne lyg panašūs, tačiau analogiškai galima pastebėti, kad yra keli įrašai, kurie išsiskiria nuo daugumos ir, skirtingai ne Andrew kreivių vizualizacijoje, čia (žr 8 Pav. ir 9 Pav.) galima pamatyti kuris požymis juos išskiria, konkrečiai - santykinis skaičių kiekis (numberCount).



8. Pav: Pandas duomenų lygiagrečių kooridinačių vizualizacija , 770 įrašų



9. Pav: Pandas duomenų lygiagrečių kooridinačių vizualizacija , 150 įrašų

4 Išvados

Grafiškai galima pamatyti kaip tarpusavyje yra pasiskirstę duomenys, kokių reikšmių yra daugiausiai, kurie atributai turi įtaką klasės atributui bei kurie metodai labiau tinka kategorijoms vizualiai atskirti el. laiškų duomenų aibėje.

Labiausiai tiko spindulinio vizualizavimo (FreeViz ir RadViz) metodai, o mažiausiai – Andrew kreivių ir lygiagrečių koordinačių metodai.