KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Modulio P160B124 „Mašininio mokymosi metodai“

Laboratorinio darbo Nr.3 ataskaita

Studentas
Oskaras Valentinavičius IFB-7

Vertina
Doc. Tomas Iešmantas

KAUNAS, 2020

TURINYS

PAVEIKSLŲ SĄRAŠAS

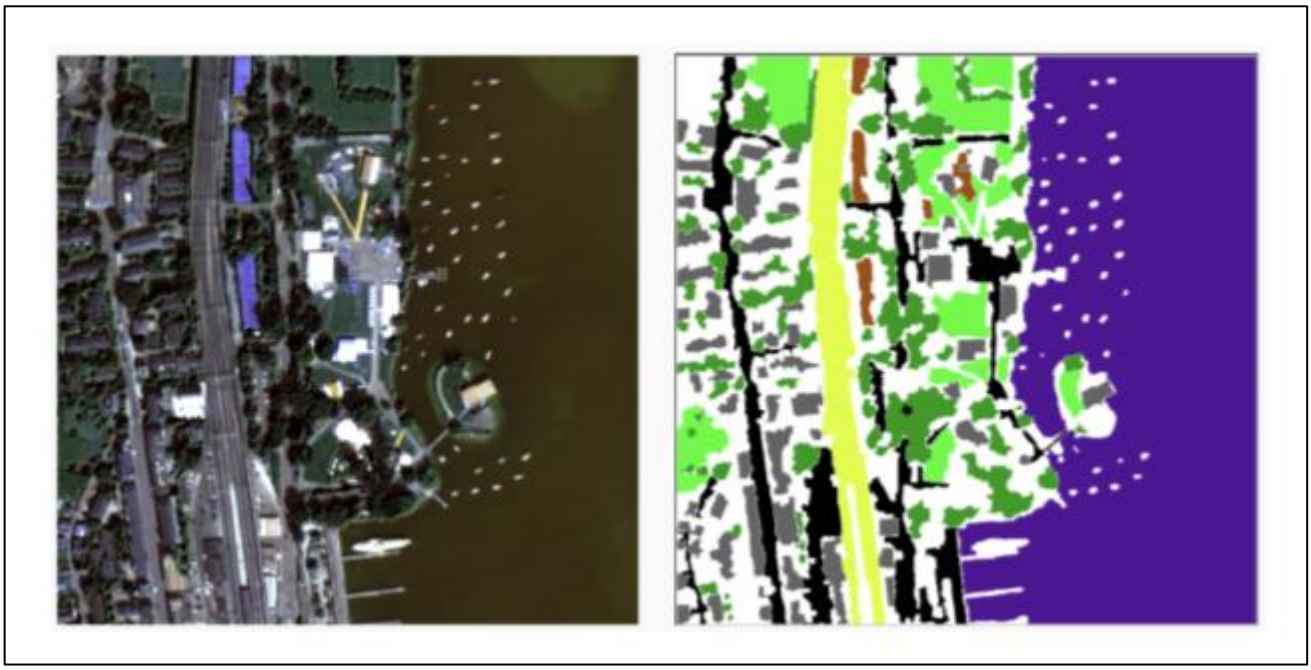
TIKSLAS

Susipažinti su pagalbinio vektoriaus mašina („Support vector machine“) ir neuroniniais tinklais („Neural networks“), norint prognozuoti reikšmių klasifikaciją. Ištirti skirtingus uždavinių parametrus, jų įtaką rezultatų tikslumui ir pritaikyti teorines žinias praktiškai. Gautus rezultatus atvaizduoti grafiškai, pakomentuoti ir įvertinti jų tikslumą.

DARBO EIGA

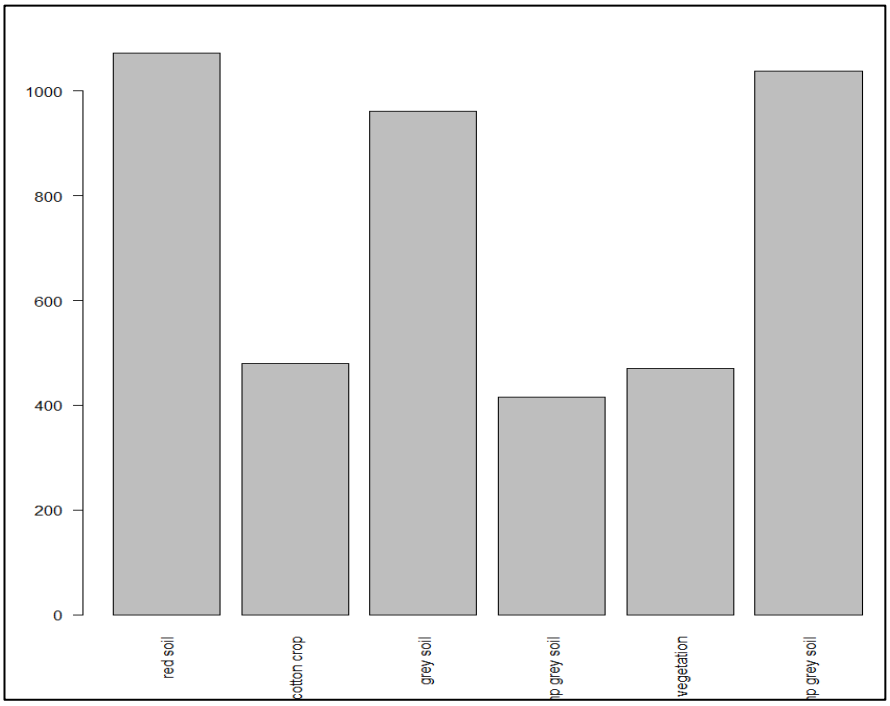
Pirma laboratorinio darbo dalis

Pirmoje laboratorinio darbo dalyje yra dirbama su duomenimis, kurie apibūdina palydovo paimtus vaizdus iš žemės. Šio darbo tikslas – naudojanti pagalbinio vektoriaus mašina („Support vector machine“) mašininio mokymosi būdu atitinkamai klasifikuoti duomenų pikselius į suteiktas kategorijas. Naudojamas duomenų rinkinys susideda iš 36 kintamųjų. Duomenys yra pateikiami „csv“ formatu. Toliau pateikiamas duomenų rinkinio vaizdas bei atliktos užduoties vaizdas didesniu formatu (pav. 1).



pav. 1 Duomenų atvaizdavimas ir jų kategorinis žemėlapis

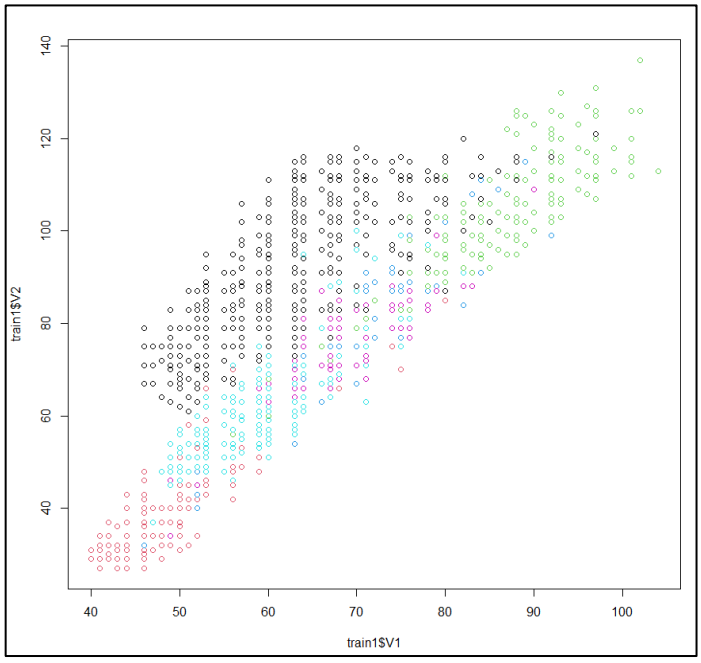
1. Suskaičiuoti ir grafiškai pateikti kiekvienos klasės dažnumus. Toliau pateikiami rezultatai (pav. 2).



pav. 2 Kiekvienos klasės dažnumas

Iš gautų rezultatų matome, jog duomenų pasiskirstymas nėra vienodas – daugiausia duomenų yra turi 1, 3 ir 6 klasės. Šios klasės turės didžiausią tikslumą prognozuojant.

2. Pasirinkti du bet kokius kintamuosius ir grafiškai atvaizduoti juos abu. X – ašyje yra atvaizduojamas „V1", o Y – ašyje „V2" kintamieji. Šie taškai yra nuspalvinami pagal atitinkamas klases (pav. 3).



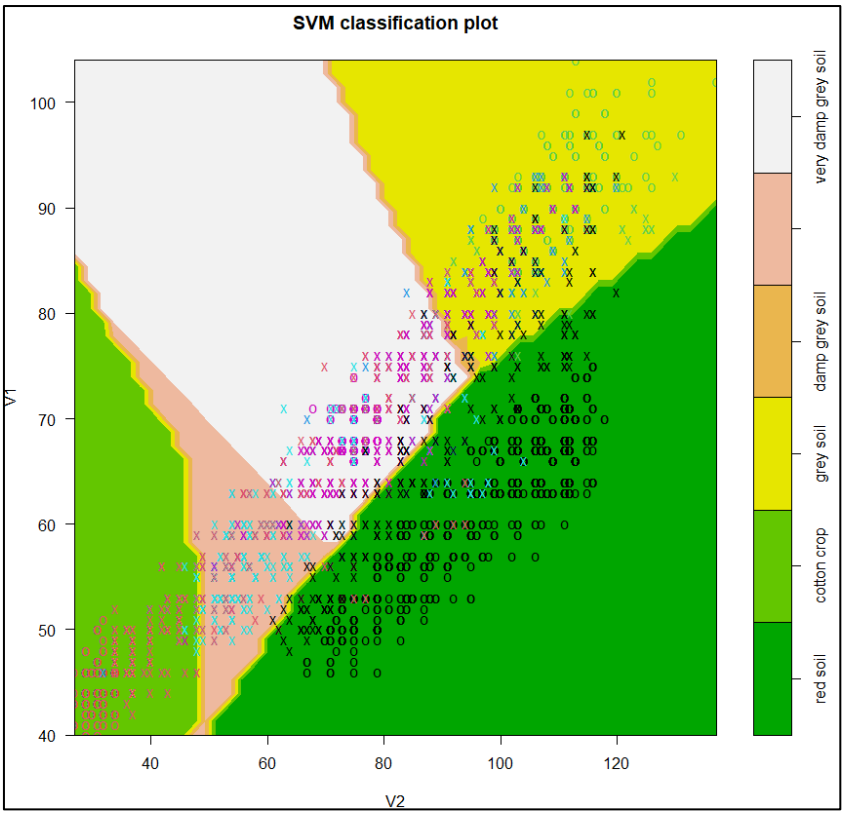
pav. 3 Dviejų parametrų atvaizdavimas

Iš gautų rezultatų galime pastebėti, kad esančios klasės tam tikra prasme „lipa“ viena ant kitos. Būtent dėl to, jei mūsų klasifikavimo metodas būtų priskirtas tik šiems dviem parametrams, jo tikslumas būtų tikrai mažas. Taip yra, nes klasės yra ganėtinai arti susigrupavusios.

3. Pasinaudojant „tune“ funkcija yra surandami atitinkamai optimalūs parametrai skirti tiesiniai „SVM“ funkcijai ir atliekamas prognozavimas. Šiai funkcijai vis dar yra naudojami du parametrai iš praeito žingsnio. Toliau pateikiami rezultatai: kryžminės patikros paklaida (pav. 4) , „SVM“ sprendimo ribos (pav. 5), prognozuotų reikšmių bendra paklaida ir sumaišties lentelė (pav. 6).

```
> summary(tune.out)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
  0.1

- best performance: 0.2263829
```

pav. 4 Optimalūs parametrai „SVM“ funkcijai



pav. 5 „SVM“ sprendimo ribos

```
> table(predicted_labels=pred, true_labels=test1$V37)
                     true_labels
predicted_labels      red soil cotton crop grey soil
  red soil                419           2          3
  cotton crop               0         184          0
  grey soil                23           5        371
  damp grey soil            0           0          0
  soil with vegetation      3          16          2
  very damp grey soil      16          17         21
                     true_labels
predicted_labels      damp grey soil soil with vegetation
  red soil                         5                    12
  cotton crop                      1                     6
  grey soil                       84                     7
  damp grey soil                   6                     2
  soil with vegetation             1                   137
  very damp grey soil            114                    73
                     true_labels
predicted_labels      very damp grey soil
  red soil                            2
  cotton crop                         2
  grey soil                          48
  damp grey soil                      2
  soil with vegetation                9
  very damp grey soil               407
> mean(test1$V37==pred)
[1] 0.762
> table1 = table(predicted=pred,true_labels=test1$V37)
> diag(table1)/colSums(table1)
           red soil          cotton crop            grey soil
         0.90889371           0.82142857           0.93450882
     damp grey soil soil with vegetation  very damp grey soil
         0.02843602           0.57805907           0.86595745
>
```

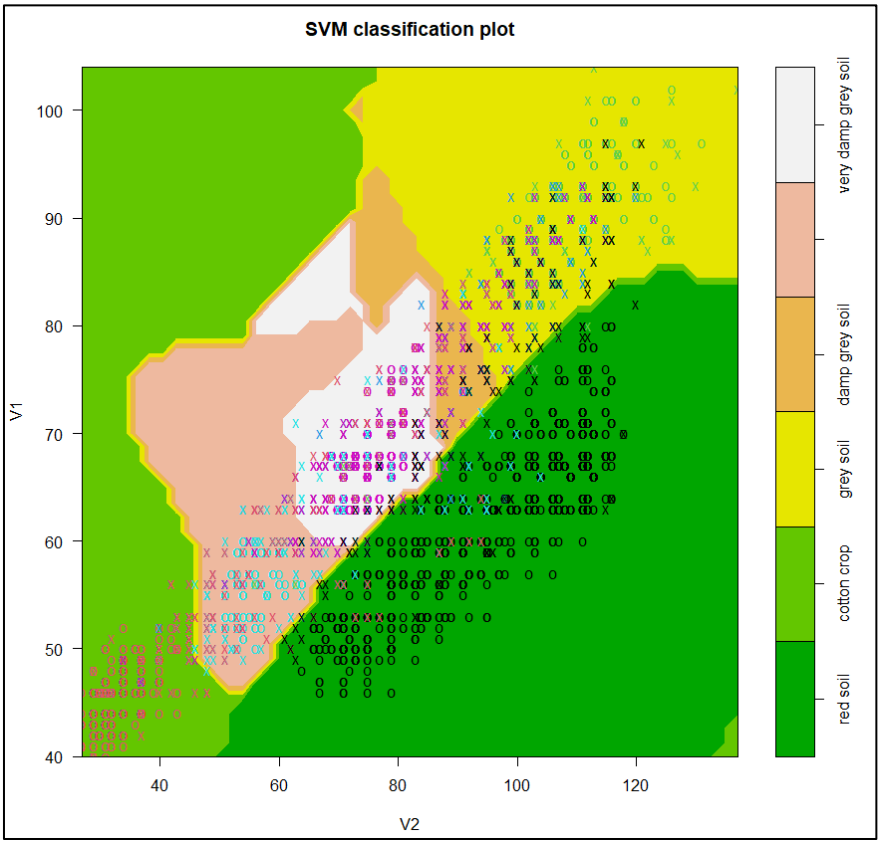
pav. 6 Bendra paklaida ir sumaišties lentelė

Iš gautų rezultatų matome, jog ne visos klasės turėjo gerą tikslumą – ketvirta ir penkta klasės turėjo mažiausius tikslumo koeficientus. Jų tikslumą nulėmė ne vien netolygus klasių pasiskirstymas dvimatėje erdvėje, bet ir tai, kad šios dvi klasės.

4. Pasinaudojant „tune“ funkcija yra surandami atitinkamai optimalūs hyper-parametrai skirti ne tiesiniai „SVM“ funkcijai, laisvai pasirenkama „kernel“ reikšmė (pasirenkama „radial“) ir atliekamas prognozavimas. Toliau pateikiami rezultatai: kryžminės patikros paklaida (pav. 7) , „SVM“ sprendimo ribos (pav. 8), prognozuotų reikšmių bendra paklaida ir sumaišties lentelė (pav. 9).

```
> summary(tune.out)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
   10     1

- best performance: 0.2128668
```

pav. 7 Optimalūs hyper-parametrai ne tiesiniai „SVM“ funkcijai



pav. 8 Ne tiesinės „SVM“ sprendimo ribos

```
> table(predicted_labels=pred, true_labels=test1$V37)
                    true_labels
predicted_labels     red soil cotton crop grey soil
  red soil                424           2          2
  cotton crop               0         184          0
  grey soil                21           3        367
  damp grey soil            2           6         12
  soil with vegetation      4          16          2
  very damp grey soil      10          13         14
                    true_labels
predicted_labels     damp grey soil soil with vegetation
  red soil                        2                    13
  cotton crop                     1                     6
  grey soil                      72                     6
  damp grey soil                 76                    12
  soil with vegetation            1                   138
  very damp grey soil            59                    62
                    true_labels
predicted_labels     very damp grey soil
  red soil                           1
  cotton crop                        2
  grey soil                         45
  damp grey soil                    43
  soil with vegetation              10
  very damp grey soil              369
> mean(test1$V37==pred)
[1] 0.779
> table1 = table(predicted=pred,true_labels=test1$V37)
> diag(table1)/colSums(table1)
          red soil          cotton crop            grey soil
         0.9197397           0.8214286           0.9244332
     damp grey soil soil with vegetation  very damp grey soil
         0.3601896           0.5822785           0.7851064
```

pav. 9 Bendra paklaida ir sumaišties lentelė su hyper-parametrais

Iš rezultatų matome, jog bendras tikslumas pagerėjo, bet tik truputi (0.017). Galime teigti, jog dirbant su dvejais parametrais, didesnio tikslumo nepasieksime, nes dvimatė erdvė nesuteikia galimybės „SVM“ algoritmui pilnai atskirti esamų klasių.

5. Pakartoti trečią ir ketvirtą žingsnį, bet naudojant visą duomenų rinkinį.
    a. Toliau pateikiami rezultatai iš trečio žingsnio - kryžminės patikros paklaida ir optimalus „cost“ parametras (pav. 10), prognozuotų reikšmių bendra paklaida ir sumaišties lentelė (pav. 11). Naudojamas tiesinis „SVM“.

```
> summary(tune.out)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
    5

- best performance: 0.1278374
```

pav. 10 Optimalūs parametrai „SVM“ funkcijai (visi duomenys)

```
> pred=predict(bestmod ,test)
> table(predicted_labels=pred, true_labels=test$V37)
                    true_labels
predicted_labels     red soil cotton crop grey soil
  red soil                454           0         3
  cotton crop               0         209         0
  grey soil                 4           0       369
  damp grey soil            0           2        22
  soil with vegetation      3          13         1
  very damp grey soil       0           0         2
                    true_labels
predicted_labels     damp grey soil soil with vegetation
  red soil                        0                     6
  cotton crop                     1                     7
  grey soil                      42                     1
  damp grey soil                 88                     5
  soil with vegetation            2                   191
  very damp grey soil            78                    27
                    true_labels
predicted_labels     very damp grey soil
  red soil                           0
  cotton crop                        1
  grey soil                         13
  damp grey soil                    45
  soil with vegetation              12
  very damp grey soil              399
> mean(test$V37==pred)
[1] 0.855
> table1 = table(predicted=pred,true_labels=test$V37)
> diag(table1)/colSums(table1)
          red soil          cotton crop            grey soil
         0.9848156           0.9330357           0.9294710
     damp grey soil soil with vegetation  very damp grey soil
         0.4170616           0.8059072           0.8489362
```

pav. 11 Bendra paklaida ir sumaišties lentelė (visi duomenys)

b. Toliau pateikiami rezultatai iš ketvirto žingsnio - kryžminės patikros paklaida ir optimalūs hyper-parametrai (pav. 12), prognozuotų reikšmių bendra paklaida ir sumaišties lentelė (pav. 13). Naudojamas ne tiesinis „SVM“ („kernel“ = „radial“).

```
> summary(tune.out)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
    1   0.5

- best performance: 0.09876457
```

pav. 12 Optimalūs hyper-parametrai ne tiesiniai „SVM“ funkcijai (visi duomenys)

```
> table(predicted_labels=pred, true_labels=test$V37)
                    true_labels
predicted_labels     red soil cotton crop grey soil
  red soil                450           0         3
  cotton crop              10         221        11
  grey soil                 0           0       370
  damp grey soil            0           0        11
  soil with vegetation      1           1         0
  very damp grey soil       0           2         2
                    true_labels
predicted_labels     damp grey soil soil with vegetation
  red soil                        0                     1
  cotton crop                     9                    10
  grey soil                      33                     0
  damp grey soil                129                     1
  soil with vegetation            2                   216
  very damp grey soil            38                     9
                    true_labels
predicted_labels     very damp grey soil
  red soil                           0
  cotton crop                        8
  grey soil                         11
  damp grey soil                    16
  soil with vegetation               9
  very damp grey soil              426
> mean(test$V37==pred)
[1] 0.906
> table1 = table(predicted=pred,true_labels=test$V37)
> diag(table1)/colSums(table1)
           red soil          cotton crop            grey soil
          0.9761388            0.9866071            0.9319899
      damp grey soil soil with vegetation  very damp grey soil
          0.6113744            0.9113924            0.9063830
>
```

pav. 13 Bendra paklaida ir sumaišties lentelė su hyper-parametrais (visi duomenys)

Iš gautų rezultatų matome, jog geriausiai prognozė atliko paskutinis modelis. Jo bendras tikslumas lygus 90.6%. Toliau pateikiamas pagalbinių vektorių kiekis (pav. 14).

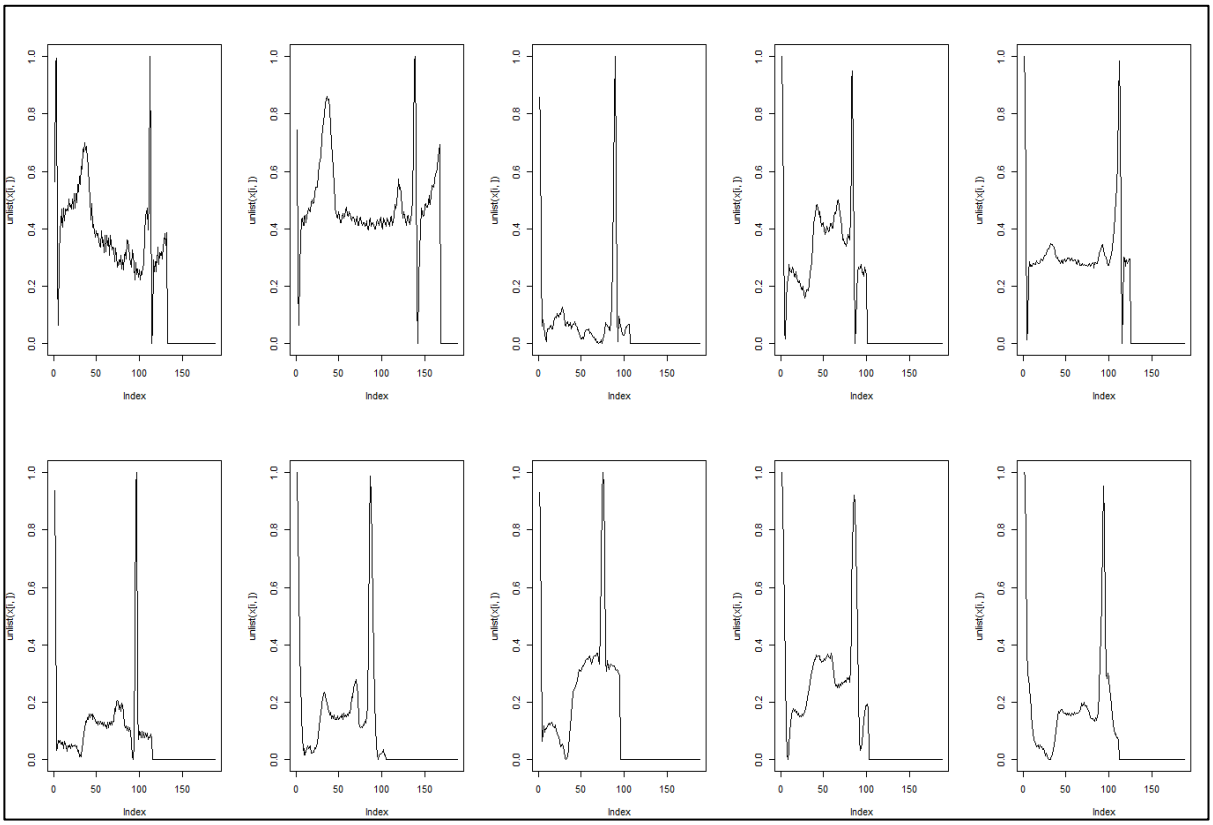
```
Call:
best.tune(method = svm, train.x = V37 ~ ., data = train,
    ranges = list(cost = c(0.1, 1), gamma = c(0.5, 1)),
    kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  2567

 ( 499 348 367 518 371 464 )


Number of Classes:  6
```

pav. 14 Pagalbinių vektorių kiekis

Antra laboratorinio darbo dalis

Antroje laboratorinio darbo dalyje yra dirbama su neuroniniais tinklais. Tam yra pasinaudojama „h2o“ biblioteka. Duomenų rinkinys apibūdina širdies dūžių elektrokardiogramos rezultatus – yra suteikiami dūžių vektoriai. Kiekvieno vektoriaus paskutinis parametras apibūdina aritmijos tipą. Darbo tikslas – pasinaudojant neuroniniu tinklu, klasifikuoti širdies dūžių vektorių aritmiją.

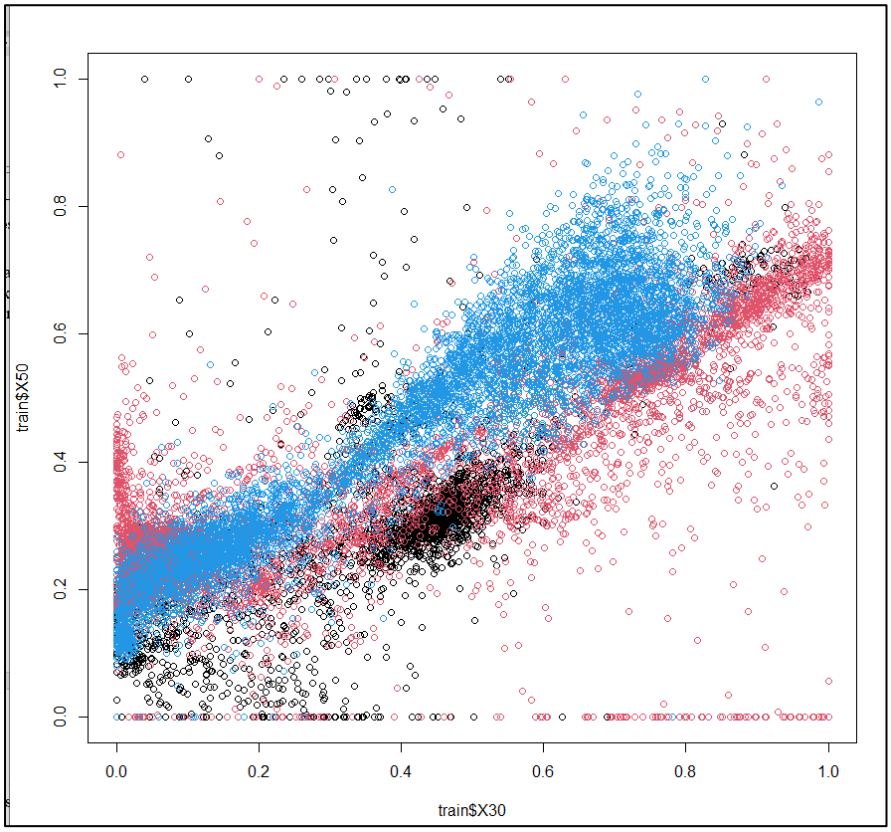
1. Atvaizduoti 10 normalių širdies smūgių vektorius (pav. 15).



pav. 15 10 normalių širdies smūgių vektoriai

Iš rezultatų galime pastebėti, jog nors ir yra atvaizduojami širdies dūžių vektoriai, kurie yra klasifikuojami kaip normalūs, vis tiek duomenų pasiskirstymas yra labai didelis. Visi dūžiai yra labai skirtingi. Tai tikrai paveiks neuroninį tinklą, todėl galimai reikės papildomų lygių.

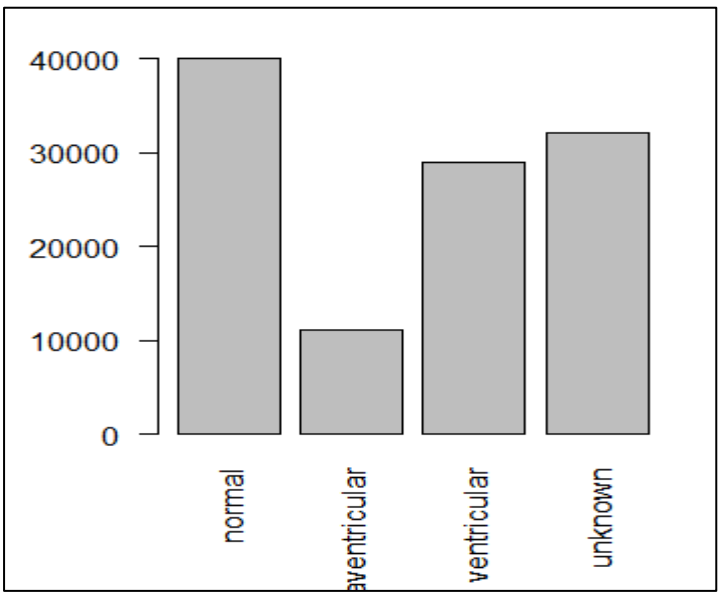
2. Toliau yra pasirenkami du skirtingi atributai ir vienas su kitu yra atvaizduojami, skirtingos spalvos vaizduoja skirtingas aritmijos klases (pav. 16).



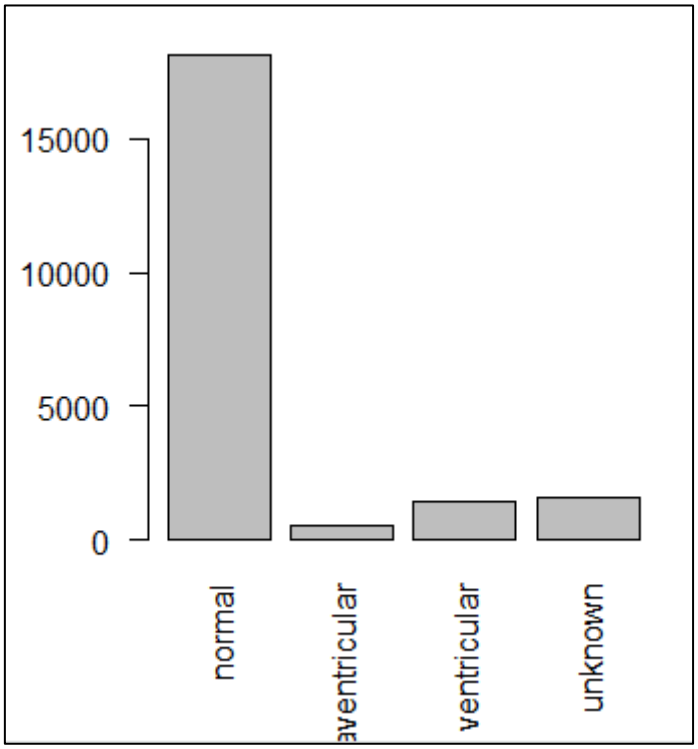
pav. 16 Dviejų atributų atvaizdavimas su klasių spalvomis

Atsižvelgiant į rezultatą, matome, jog atskirti klases pagal du parametrus yra labai sudėtinga.

3. Atvaizduoti aritmijos klasių pasiskirstymus iš treniravimo ir testavimo duomenų rinkinių (pav. 17 – 18).



pav. 17 Testavimo duomenų aritmijos klasių pasiskirstymas



pav. 18 Treniravimo duomenų aritmijos klasių pasiskirstymas

Iš gautų grafikų matome, jog mažiausiai duomenų turi „supraventricular" aritmijos klasė. Pagal tai galima manyti, kad ši klasė turės didžiausia klaidų kiekį ir mažiausią tikslumą.

4. Pateikti gautus rezultatus iš sukurtų neuroninių tinklų prognozių, kuriuose skiriasi epochų kiekiai, aktyvacijos funkcijos ir reguliarizacijos.
    a. Toliau pateikiami rezultatai iš neuroninio tinklo, kuris turi 100 epochų, naudoja „Tanh" aktyvacijos funkcija ir turi viena paslėptą sluoksnį su 10 neuronų (pav. 19).

```
> mean(prediction$predict==test$arrhythmia)
[1] 0.9048276
> confMatrix=table(predicted_labels=prediction$predict,
+                  true_labels=test$arrhythmia)
> diag(confMatrix)/colSums(confMatrix)
          normal supraventricular      ventricular
      0.90495115       0.71043165       0.01035912
         unknown
      0.02363184
```

pav. 19 Pirmo bandymo rezultatai

b. Toliau pateikiami rezultatai iš neuroninio tinklo, kuris turi 1000 epochų, naudoja „Tanh“ aktyvacijos funkcija ir turi du paslėptus sluoksnius po 10 neuronų ir naudojami yra L1 = 0, L2 = 0.01 reguliarizacijos parametrai (pav. 20).

```
> mean(prediction$predict==test$arrhythmia)*100
[1] 91.03042
> confMatrix=table(predicted_labels=prediction$predict,
+                  true_labels=test$arrhythmia)
> diag(confMatrix)/colSums(confMatrix)*100
          normal supraventricular      ventricular
     91.2789093       61.8705036        0.9668508
         unknown
      1.4303483
```

pav. 20 Antro bandymo rezultatai

c. Toliau pateikiami rezultatai iš neuroninio tinklo, kuris turi 100 epochų, naudoja „Rectifier“ aktyvacijos funkcija ir turi vieną paslėptą sluoksnį su 10 neuronų ir yra naudojami L1 = 0, L2 = 0.01 reguliarizacijos parametrai (pav. 21).

```
> mean(prediction$predict==test$arrhythmia)*100
[1] 93.89756
> confMatrix=table(predicted_labels=prediction$predict,
+                  true_labels=test$arrhythmia)
> diag(confMatrix)/colSums(confMatrix)*100
          normal supraventricular      ventricular
     96.5170834       57.1942446        2.7624309
         unknown
      0.3731343
```

pav. 21 Trečio bandymo rezultatai

d. Toliau pateikiami rezultatai iš neuroninio tinklo, kuris turi 1000 epochų, naudoja „Rectifier“ aktyvacijos funkcija ir turi du paslėptus sluoksnius po 10 neuronų ir yra naudojami L1 = 0, L2 = 0.01 reguliarizacijos parametrai (pav. 22).

```
> mean(prediction$predict==test$arrhythmia)*100
[1] 93.23025
> confMatrix=table(predicted_labels=prediction$predict,
+                  true_labels=test$arrhythmia)
> diag(confMatrix)/colSums(confMatrix)*100
          normal supraventricular      ventricular
     93.7627643       62.2302158        0.6906077
         unknown
      0.8706468
```

pav. 22 Ketvirto bandymo rezultatai

Iš rezultatų matome, kad geriausią rezultatą davė trečias bandymas. Mažiausia tikslumą turėjo „unknown“ klasė šiame bandyme, nors pačioje pradžioje, dėl klasių pasiskirstymo, atrodė, kad tai bus „supraventricular“ klasė.

IŠVADOS

Atlikus laboratorinį darbą, buvo pagilintos teorinės žinios apie neuroninius tinklus ir pagalbinę vektoriaus mašiną. Žinios buvo užtvirtintos praktiškai.