Vilniaus Universitetas

Tiesioginio sklidimo DNT naudojant sistemą WEKA

Dirbtinio intelekto pagrindai

3 užduotis

Darbą atliko:

Dovydas Martinkus

Duomenų Mokslas 4 kursas 1gr.

Vilnius, 2022

**Turinys**

[1 Tikslas ir uždaviniai 3](#_Toc114680263)

[2 Duomenys 4](#_Toc114680264)

[3 Užduoties ataskaita 5](#_Toc114680265)

[4 Išvados 14](#_Toc114680266)

# Tikslas ir uždaviniai

Tikslas: Išmokyti neuroninį tinklą teisingai klasifikuoti duomenis naudojant sistemą WEKA.

Uždaviniai:

Analizuojamų duomenų paruošimas ir aprašymas.

Užduočių sekų sudarymas ir jų vaizdinis pateikimas su komentarais kam skirta kiekviena naudojama komponentė.

Neuroninio tinklo parametrų parinkimas: optimalių paslėptų neuronų skaičiaus, mokymo greičio parametro bei momentum reikšmių radimas pateikiant klasifikavimo tikslumo įverčius visiems tirtiems atvejams.

Naujų duomenų klasifikavimas naudojant išsaugotą tinklo modelį.

Neuronų išėjimo reikšmių perskaičiavimas „rankiniu“ būdu pasirinktoje programoje. Išvadų apie rezultatų sutapimą padarymas.

# Duomenys

Užduotyje naudojamą irisų duomenų aibę sudaro 150 stebėjimų, iš kurių kiekvienas turi po 4 skaitinius požymius ir klasę, kuriai priklauso. Šiuose duomenyse yra trys klasės: Setosa, Versicolor ir Virginica. Duomenų aibėje yra po 50 kiekvienos klasės stebėjimų.

Analizei naudotas *.arff* failas su šia duomenų aibe, kuris yra įrašomas į kompiuterį įdiegus sistemą WEKA.

Papildomai prieiga prie duomenų rinkinio per internetą: <https://archive.ics.uci.edu/ml/datasets/iris>.

# Užduoties ataskaita

Duomenų pertvarkymas atliktas naudojant programavimo kalbą „Python“. Duomenys padalinti į mokymo ir testavimo aibes santykiu 80-20 abiejose aibėse išlaikant tokį patį skirtingų klasių stebėjimų skaičiaus santykį (mokymo aibėje tokiu būdu paliekant po 40 kiekvienos klasės stebėjimų). Mokymo ir testavimo aibės atitinkamai išsaugotos į *iris\_train\_test.arff* ir *iris\_new.arff* failus:

**from** scipy.io **import** arff **as** sciarff

**import** pandas **as** pd

**from** sklearn.model\_selection **import** train\_test\_split

**from** collections **import** Counter

data **=** sciarff**.**loadarff('iris.arff')

df **=** pd**.**DataFrame(data[0])

df**.**columns **=** ["sepal\_length","sepal\_width","petal\_length","petal\_width","class"]

df **=** df**.**replace({b'Iris-setosa': 'setosa', b'Iris-virginica': 'virginica', b'Iris-versicolor' : 'versicolor'})

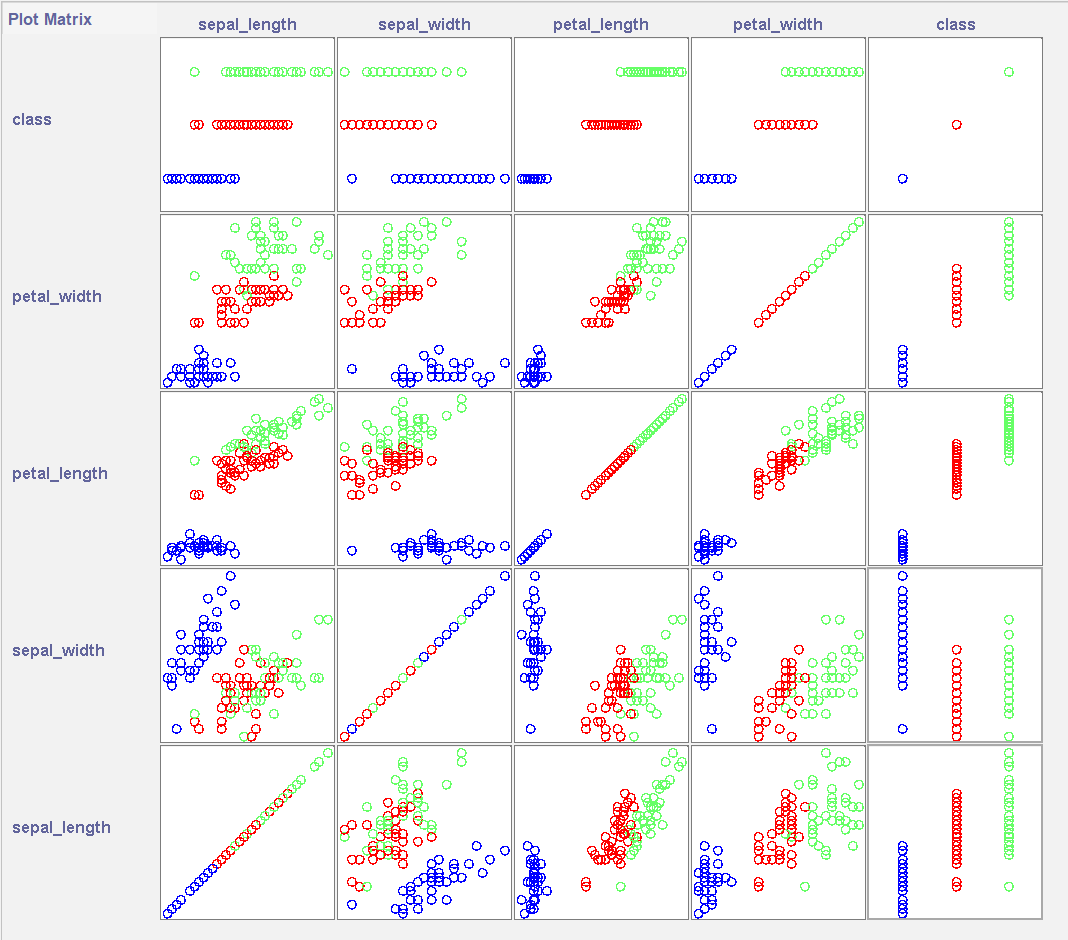
train, test **=** train\_test\_split(df,train\_size**=**0.8,test\_size**=**0.2,stratify**=**df["class"])

**from** pandas2arff **import** pandas2arff

pandas2arff(train, "iris\_train\_test.arff", wekaname**=**"iris\_train\_test")

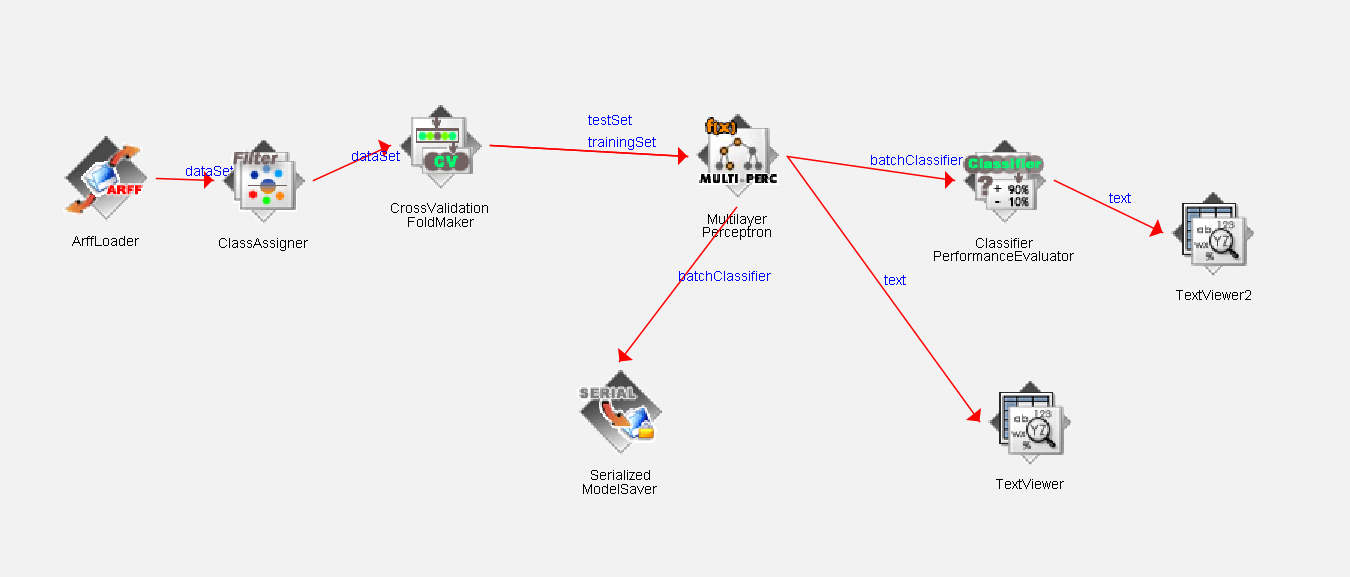
pandas2arff(test, "iris\_new.arff", wekaname**=**"iris\_new")

Žemiau pateikta skaitinių kintamųjų sklaidos diagramų matrica testavimo aibėje esantiems stebėjimams:



1 pav. Duomenų aibėje esančių stebėjimų skaitinių požymių sklaidos diagramų matrica

Žemiau pavaizduota WEKA sistemoje sudaryta užduočių seka, skirta surasti geriausius neuroninio tinklo parametrų rinkinius, klasifikavimo vertinimui vietoje testavimo aibės naudojant kryžminę validaciją:



2 pav. Pirmos užduočių sekos vaizdas

Užduočių sekoje naudojami šie komponentai:

„ClassAssigner“ - skirtas sekančioms komponentėms pažymėti, kuris duomenų stulpelis skirtas klasei.

„CrossValidationFoldMaker“ – atlieka kryžminę validaciją.

„TextViewer“ ir „TextViewer2“ – atitinkamai skirti peržiūrėti gautus neuroninio tinklo svorius ir klasifikavimo rezultatų įvertinimus.

„SerializedModelSaver“ – išsaugo gautą modelį.

„ClassifierPerfomanceEvaluator“ – įvertina klasifikavimo rezultatus.

„MultilayerPerceptron“ – sudaro ir apmoko daugiasluoksnį perceptroną su pasirinktais parametrais.

Naudojant šią seką keisti mokymosi greičio, momentum ir paslėptų neuronų skaičiaus parametrai, siekiant surasti parametrų reikšmių rinkinį, su kuriais gaunami geriausi klasifikavimo rezultatai. Kaip matome iš žemiau pateiktų paveikslėlių, geriausi klasifikavimo rezultatai (96,6 % tikslumas) gaunami naudojant 2 paslėptų neuronų sluoksnius, 0,1 mokymo greitį ir 0,1 momentum reikšmę (taip pat ir naudojant 0,3 momentum reikšmę):

Paveikslėlis, kuriame yra žinutė

Automatiškai sugeneruotas aprašymas

3 pav. Klasifikavimo rezultatai naudojant kryžminę validaciją (2 paslėpti sluoksniai, 0,1 mokymo greitis, 0,1 momentum)

Paveikslėlis, kuriame yra žinutė

Automatiškai sugeneruotas aprašymas

4 pav. Klasifikavimo rezultatai naudojant kryžminę validaciją (2 paslėpti sluoksniai, 0,3 mokymo greitis, 0,3 momentum)

Paveikslėlis, kuriame yra žinutė

Automatiškai sugeneruotas aprašymas

5 pav. Klasifikavimo rezultatai naudojant kryžminę validaciją (3 paslėpti sluoksniai, 0,1 mokymo greitis, 0,1 momentum)

Paveikslėlis, kuriame yra žinutė

Automatiškai sugeneruotas aprašymas

6 pav. Klasifikavimo rezultatai naudojant kryžminę validaciją (3 paslėpti sluoksniai, 0,3 mokymo greitis, 0,3 momentum)

Paveikslėlis, kuriame yra žinutė

Automatiškai sugeneruotas aprašymas

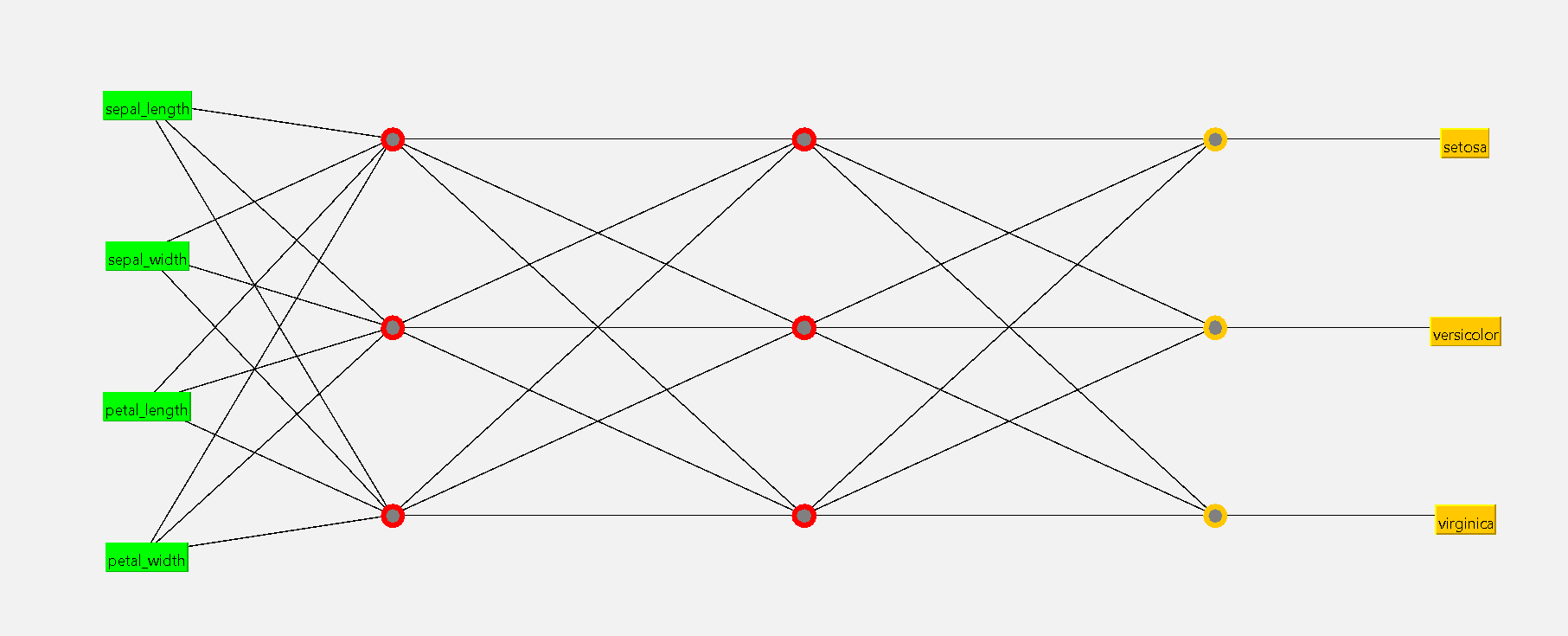
7 pav. Klasifikavimo rezultatai naudojant kryžminę validaciją (2 paslėpti sluoksniai, 0,3 mokymo greitis, 0,1 momentum)

Paveikslėlis, kuriame yra žinutė

Automatiškai sugeneruotas aprašymas

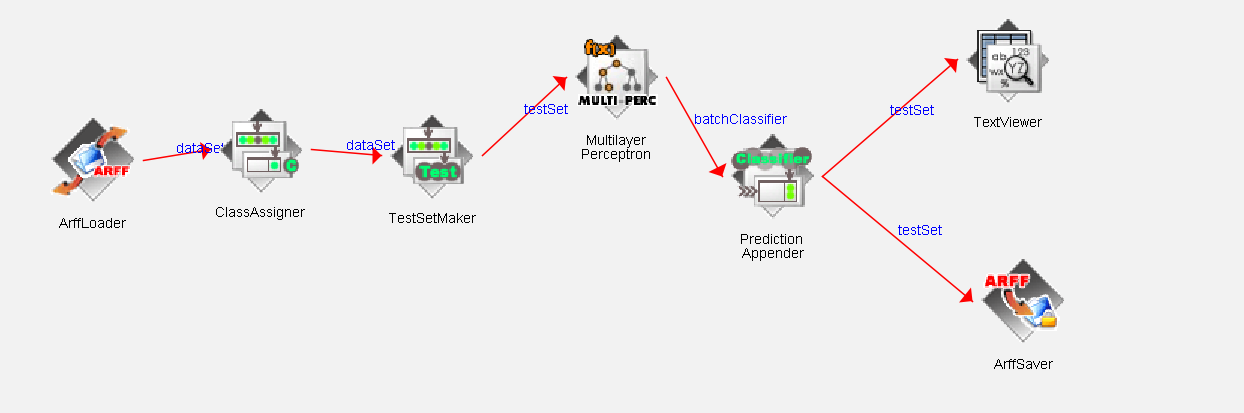
8 pav. Klasifikavimo rezultatai naudojant kryžminę validaciją (2 paslėpti sluoksniai, 0,1 mokymo greitis, 0,3 momentum)

Žemiau pateiktas geriausius rezultatus pasiekusio neuroninio tinklo grafinis vaizdas:



9 pav. Sudaryto dirbtinio neuroninio tinklo vaizdas

Sudaryta ir žemiau pavaizduota WEKA užduočių seka, skirta klasifikuoti prieš tai nematytus duomenis naudojant praeitoje užduočių sekoje išsaugotą modelį:



10 pav. Antros užduočių sekos vaizdas

Prieš tai nenaudoti užduočių sekų komponentai:

„TestSetMaker“ – paverčia duomenų aibę į testavimo (sekančioms komponentėms).

„PredictionAppender“ – skirtas papildyti duomenis prognozuotos klasės stulpeliu.

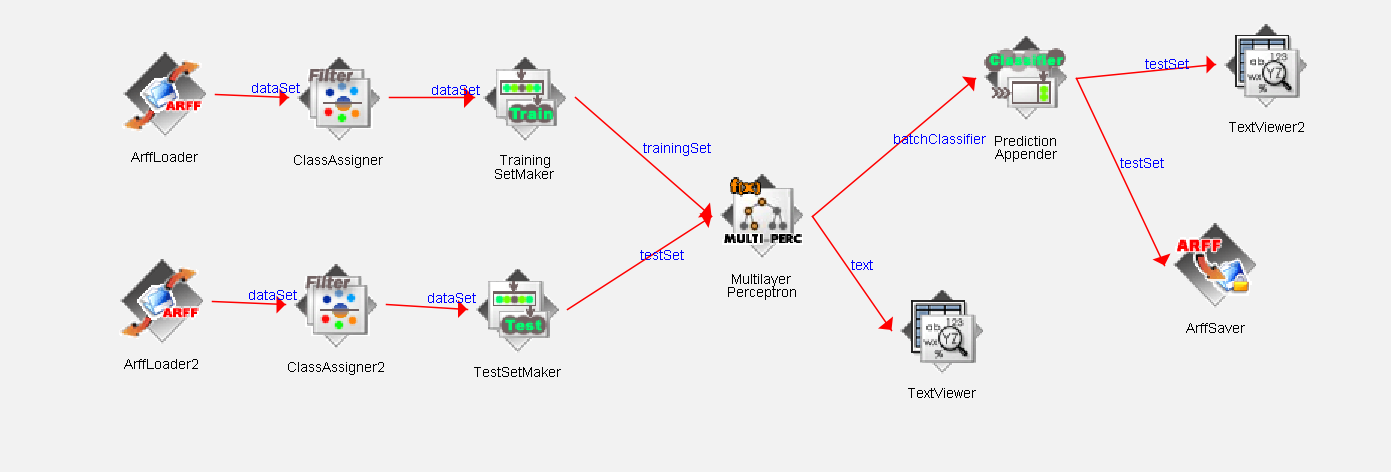
„ArffSaver“ – išsaugo duomenis, papildytus prognozuotos klasės stulpeliu.

Lentelėje palyginimui pateiktos tikros klasės ir prognozuotos klasės, gautos naudojant išsaugotą modelį (1 lentelė). Matome, kad visais atvejais tikros ir modelio prognozuotos klasės sutapo (turimas 100% tikslumas).

1 lentelė Tikros ir prognozuotos klasės testavimo aibei naudojant išsaugotą modelį

|  |  |
| --- | --- |
| Tikra klasė | Prognozuota klasė |
| setosa | setosa |
| setosa | setosa |
| versicolor | versicolor |
| versicolor | versicolor |
| versicolor | versicolor |
| versicolor | versicolor |
| setosa | setosa |
| virginica | virginica |
| setosa | setosa |
| versicolor | versicolor |
| virginica | virginica |
| versicolor | versicolor |
| versicolor | versicolor |
| virginica | virginica |
| setosa | setosa |
| versicolor | versicolor |
| setosa | setosa |
| virginica | virginica |
| virginica | virginica |
| virginica | virginica |
| setosa | setosa |
| virginica | virginica |
| setosa | setosa |
| setosa | setosa |
| virginica | virginica |
| virginica | virginica |
| virginica | virginica |
| versicolor | versicolor |
| setosa | setosa |
| versicolor | versicolor |

Žemiau pateiktas užduočių sekos, sudarytos siekiant apmokyti modelį naudojant visą testavimo aibę ir klasifikuoti stebėjimus testavimo aibėje, vaizdas:



11 pav. Trečios užduočių sekos vaizdas

Šioje sekoje yra tik 1 prieš tai nenaudota komponentė:

„TrainingSetMaker“ – iš duomenų sudaro mokymo aibę (sekančioms komponentėms).

Pasirinkta sudaryti neuroninį tinklą su 1 paslėptų sluoksniu, kurį sudaro 5 neuronai (naudojant numatytas mokymosi greičio ir momentum reikšmes, kurios lygios atitinkamai 0,3 ir 0,2). Apmokyto modelio tikslumas testavimo aibėje gautas lygus 96,6 %. Į išvesties failą išsaugotos modeliu gautos spėjamos klasės testavimo duomenims kartu su tikimybėmis priklausyti kiekvienai klasei.

Žemiau pateikti (2 ir 3 lentelės) gauti neuroninio tinklo neuronų svoriai (atskirai paslėptam ir išvesties sluoksniams):

2 lentelė Išvesties sluoksnio svoriai skirtingiems neuronams

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
| Poslinkis | 1,032 | -3,560 | -3,690 |
| 3 | 3,657 | -0,823 | -7,646 |
| 4 | -3,694 | 4,242 | 0,244 |
| 5 | -1,770 | -6,863 | 5,331 |
| 6 | -1,841 | -7,726 | 5,833 |
| 7 | -5,342 | 7,392 | 0,609 |

3 lentelė Paslėpto sluoksnio svoriai skirtingiems neuronams

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 3 | 4 | 5 | 6 | 7 |
| Poslinkis | 2,701 | 1,898 | -5,584 | -6,084 | 2,482 |
| sepal\_length | 1,514 | 1,150 | -2,324 | -2,725 | 1,313 |
| sepal\_width | 0,983 | -1,943 | -1,814 | -2,089 | -2,547 |
| petal\_length | -5,153 | 2,916 | 8,838 | 9,825 | 3,782 |
| petal\_width | -2,476 | 2,514 | 4,868 | 5,145 | 3,393 |

Naudojant šiuos išsaugotus svorius „Excel“ programa rankiniu būtu atkartotas WEKA atliktas klasių prognozių gavimas.

Duomenys, esantys testavimo aibėje perkelti į „Excel“, kartu su išsaugotais svoriais. Testavimo aibės duomenys normuoti intervale [-1 ; 1].

Skaičiavimai atlikti naudojant tokią tvarką:

Pirmiausia kiekvienam iš 5 paslėpto sluoksnio neuronų susumuotos duomenų įėjimo vektorių (4 įvestys) ir paslėptų neuronų svorių vektorių (4 svoriai ir poslinkis) sandaugos.

Šios reikšmės pateiktos sigmoidinei funkcijai. Taip gautos paslėpto sluoksnio išvesties reikšmės.

Šios 5 reikšmės naudotos kaip įvestys kiekvienam iš 3 neuronų, esančių paskutiniame (išvesties) sluoksnyje.

Galutinės priklausimo klasei tikimybės gautos pateikus 3 praėjusiame žingsnyje gautas reikšmes softmax funkcijai.

Klasės priskyrimas atliktas pasirenkant klasę, kuriai gauta didžiausia tikimybe.

Kaip matome iš žemiau esančios lentelės (4 lentelė), „Excel“ atliktais skaičiavimais gautos prognozuotos klasės visiems testavimo aibės elementams sutapo su rezultatais, gautais naudojant WEKA sistemą:

4 lentelė Excel atliktais skaičiavimais ir WEKA gautų klasių palyginimas

|  |  |
| --- | --- |
| Skaičiavimais gauta klasė | WEKA gauta klasė |
| setosa | setosa |
| setosa | setosa |
| versicolor | versicolor |
| versicolor | versicolor |
| versicolor | versicolor |
| versicolor | versicolor |
| setosa | setosa |
| virginica | virginica |
| setosa | setosa |
| versicolor | versicolor |
| virginica | virginica |
| versicolor | versicolor |
| versicolor | versicolor |
| virginica | virginica |
| setosa | setosa |
| versicolor | versicolor |
| setosa | setosa |
| virginica | virginica |
| virginica | virginica |
| virginica | virginica |
| setosa | setosa |
| versicolor | versicolor |
| setosa | setosa |
| setosa | setosa |
| virginica | virginica |
| virginica | virginica |
| virginica | virginica |
| versicolor | versicolor |
| setosa | setosa |
| versicolor | versicolor |

# Išvados

Vertinimui naudojant kryžminę validaciją geriausi klasifikavimo rezultatai (96,6 % tikslumas) buvo gauti naudojant 2 paslėptų neuronų sluoksnius, 0,1 mokymo greitį ir 0,1 momentum reikšmę (taip pat ir naudojant 0,3 momentum reikšmę). Šis modelis teisingai klasifikavo visus stebėjimus testavimo aibėje.

Keičiant mokymosi greičio ir momentum reikšmes, pastebėta, kad daugeliu atveju naudojant mažesnės mokymosi greičio ir momentum reikšmes gauti geresni rezultatai. Taip pat pastebėtas faktas, kad naudojant didesnį neuronų sluoksnių skaičių galima gauti prastesnius rezultatus, negu naudojant paprastesnį dirbtinį neuroninį tinklą (kaip tik toks atvejis ir gautas padidinus paslėptų sluoksnių skaičių nuo 2 iki 3).

„Excel“ programoje atliktais skaičiavimais gautos klasės sutapo su WEKA sistemos gautomis klasėmis. Šis rezultatas leidžia teigti, kad teorinis dirbtinio neuroninio tinklo supratimas sutapo su modeliu, kuris buvo implementuotas naudojant WEKA sistemą.