VILNIAUS UNIVERSITETSS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
PROGRAMŲ SISTEMŲ BAKALAURO STUDIJŲ PROGRAMA

Tiesioginio sklidimo DNT modeliavimas naudojant sistemą WEKA

Feedforward Neural Network Model Creation in WEKA

Laboratorinio darbo ataskaita

Atliko: Armintas Pakenis

Darbo vadovas: prof. dr. Olga Kurasova

Vilnius – 2023

TURINYS

1. Užduoties tikslas

Užduoties tikslas — išmokyti neuroninį tinklą teisingai klasifikuoti duomenis naudojant sistemą WEKA. Išsaugotus WEKA procesus ir gautus programos rezultatus galima rasti GitHub repositorijoje: https://github.com/ArmintasP/Computational-intelligence/tree/main/Lab3.

1.1. Duomenys

Naudotas irisų duomenų rinkinys: https://archive.ics.uci.edu/ml/datasets/iris.

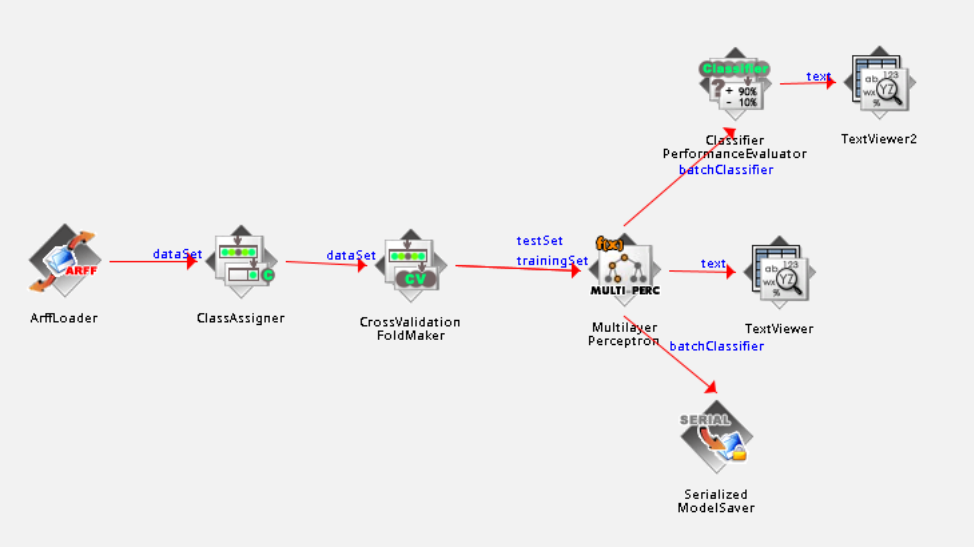
Irisų duomenų rinkinys buvo išskaidytas į du rinkinius. Vieną rinkinį sudarė 120 įrašų jis buvo naudojamas modeliui mokyti ir validuoti, tad jį naudojant buvo vykdoma kryžminė patikra. Kitą rinkinį sudarė likę 30 įrašų (po 10 iš kiekvienos klasės), kuris buvo naudojamas tik modeliui validuoti.

2. Užduočių sekų modeliavimas

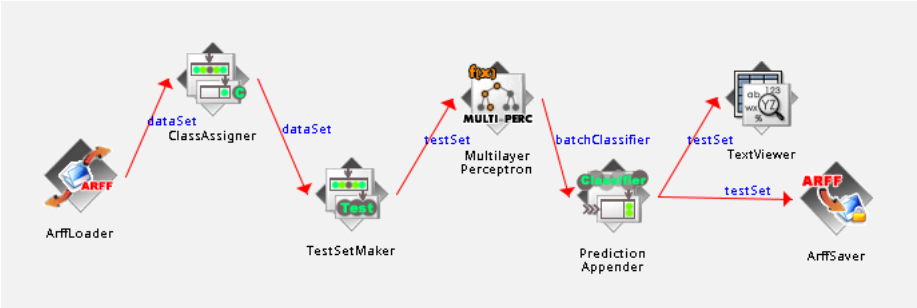
1 paveikslėlyje matoma užduoties seka naudoja 120 įrašų duomenų rinkinį. *ArffLoader* komponentas užkrauna duomenų rinkinį, *ClassAssigner* nustato, kelinta įeitis bus laikoma klasės žyme. *CrossValidation FoldMaker* suskaido duomenų rinkinį į N poabių (darbe buvo naudojama 10 poaibių), kur N-1 poaibių bus naudojama mokymui, o likęs vienas — validavimui. *Multilayer Perceptron* — neuroninis tinklas, iš kurio rezultatai keliauja į *Classifier Performance Evaluator*, kuris suskaičiuoja statistikas ir sukuria atspausdinimo lentelę. Rezultatams peržiūrėti naudojamas komponentas *TextViewer*, o neuroninio tinklo modeliui ir rastiems svoriams išsaugoti naudojamas *SerializedModelSaver*.

2 paveikslėlyje pavaizduota seka skirta tik naujiems duomenims klasifikuoti. Komponentas *TestSetMaker* duomenų rinkinį paruošia testavimui, *PredictionAppender* prie pradinio rinkinio prideda gautas modelio prognozes, o *ArffSaver* naudojamas gautam rezultatui išsaugoti.

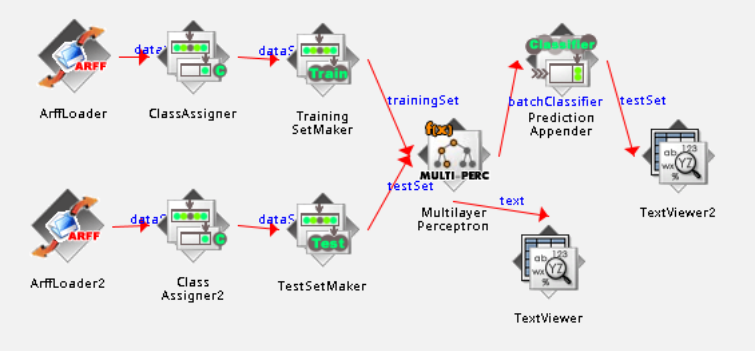
3 paveikslėlyje pavaizduota seka naudoja 120 įrašų modeliui treniruoti ir 30 įrašų iš kito failo modeliui testuoti.



1 pav. Užduočių seka su treniravimu ir validavimu naudojant duomenų failą iris_train_test.arff ir kryžminę patikrą



2 pav. Užduočių seka su validavimu naudojant duomenų failą iris_new.arff

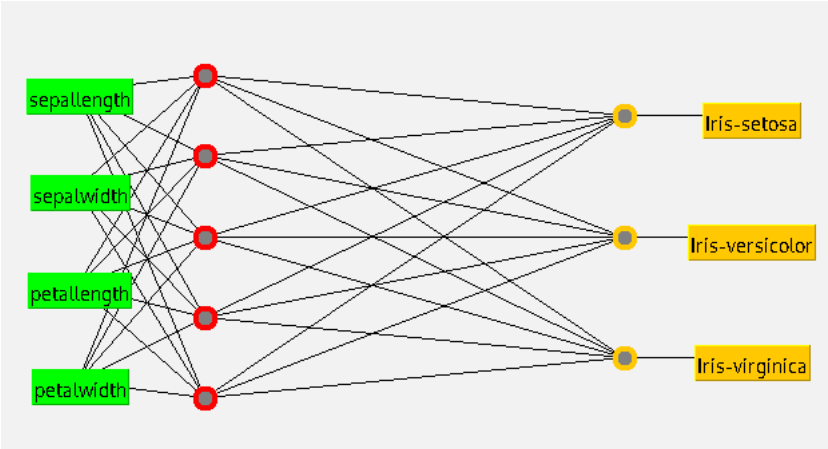


3 pav. Užduoties seka su mokymu ir validavimu

2.1. Pirmos užduočių sekos neuroninis tinklas

Iš 1 lentelės duomenų matyti, kad paslėptų neuronų skaičius ir inercija turi labai nereikšmingą įtaką modelio tikslumui. Taip dalinai yra todėl, kad duomenų rinkinys labai paprastas ir mažas, todėl užtenka ir parpastesnio neuroninio tinklo norint pasiekti aukštą tikslumą. Labiausiai tikslumas priklauso nuo mokymosi greičio, tą galime matyti iš tikslumo reikšmės nukritimo nuo 96,67 % iki 66,67 %, kai mokymosi greitis pakeičiamas iš 0,05 į 0,001.

1 lentelės pirmos eilutės parametrai buvo naudojami užkraunant neuroninio tinklo modelį antroje užduoties sekoje. Neurinio tinklo vaizdas pateiktas 4 paveikslėlyje.



4 pav. Neuroninio tinklo vaizdas

2.2. Klasifikavimo rezultatai iš WEKA sistemos

Antrojoje sekoje naudjant pirmos sekos neuroninio tinklo modelį su nauju rinkiniu (*iris_new.arff*) visos klasifikavimo prognozės sutampa su trokštomomis klasėmis, išskyrus vieną atvejį, kuris paršykintas 2 lentelėje. Todėl galima teigti, kad tinklas klasifikavo labai gerai nematytą duomenų rinkinį, kadangi tikslumas $\frac{29}{30} * 100\% \approx 96{,}67\%$ sutampa su mokymosi geriausiu mokymosi ir testavimo tikslumu, kuris pateiktas 1 lentelėje. Taip pat iš 5 pav. matyti, kad kiekvienos klasės požymiai klasterizuojasi su toje pačioje klasėje.

1 lentelė. Dirbtinio neuroninio tinklo tikslumo priklausomybė nuo parametrų, epochų skaičius = 500

Paslėptų neuronų sk.	**Mokymosi greitis**	**Inercija**	**Tikslumas**
5	0,050	0,40	96,67 %
1	0,050	0,40	95,00 %
0	0,050	0,40	94,17 %
5	0,200	0,40	95,83%
5	0,001	0,40	66,67 %
5	0,050	0,10	96,67 %
5	0,050	0,80	96,67 %
5	0,050	0,01	95,83%
5	0,050	0,99	95,83%

2 lentelė. Antros užduočių sekos prognozės naujam rinkiniui

sepallength	**sepalwidth**	**petallength**	**petalwidth**	**Trokštama klasė**	**Prognozė**
5,1	3,5	1,4	0,2	Iris-setosa	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa	Iris-setosa
5	3,6	1,4	0,2	Iris-setosa	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa	Iris-setosa
5	3,4	1,5	0,2	Iris-setosa	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa	Iris-setosa
4,9	3,1	1,5	0,1	Iris-setosa	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor	Iris-versicolor
6,9	3,1	4,9	1,5	Iris-versicolor	Iris-versicolor
5,5	2,3	4	1,3	Iris-versicolor	Iris-versicolor
6,5	2,8	4,6	1,5	Iris-versicolor	Iris-versicolor
5,7	2,8	4,5	1,3	Iris-versicolor	Iris-versicolor
6,3	3,3	4,7	1,6	Iris-versicolor	Iris-versicolor
4,9	2,4	3,3	1	Iris-versicolor	Iris-versicolor
6,6	2,9	4,6	1,3	Iris-versicolor	Iris-versicolor
5,2	2,7	3,9	1,4	Iris-versicolor	Iris-versicolor
6,3	3,3	6	2,5	Iris-virginica	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica	Iris-virginica
7,1	3	5,9	2,1	Iris-virginica	Iris-virginica
6,3	2,9	5,6	1,8	Iris-virginica	Iris-virginica
6,5	3	5,8	2,2	Iris-virginica	Iris-virginica
7,6	3	6,6	2,1	Iris-virginica	Iris-virginica
4,9	2,5	4,5	1,7	**Iris-virginica**	**Iris-versicolor**
7,3	2,9	6,3	1,8	Iris-virginica	Iris-virginica
6,7	2,5	5,8	1,8	Iris-virginica	Iris-virginica
7,2	3,6	6,1	2,5	Iris-virginica	Iris-virginica

5 pav. Duomenų požymių porų vaizdai. Tamsiai mėlyna — Iris-setosa, raudona — Iris-versicolor, žydra — Iris-virginica

2.3. Trečios sekos modelio parametrais klasifikavimas skaičiuoklėje

Naudojant atviro kodo skaičiuoklę *LibreOffice Calc* , trečioje sekoje gautus pirmo paslėpto sluoksnio neuronų svorius (žr. 3 lentelę) ir išeities trijų neuronų svorius žr. 4 lentelę buvo sukurtas prognozavimas skaičiuoklėje. Skaičiuoklės failą galima rasti repozitorijoje.

Neuroninis tinklas (tik klasifikavimo prognozei) konstruojamas buvo taip: kiekvieno įrašo požymiai buvo padauginami iš atitinkamų paslėpto tinklo neurono svorių, kur daugybos rezultatai buvo sudedami ir dar pridedamas poslinkis. Gauta reikšmė buvo įstatoma į sigmoidinę aktyvacijos funkciją. Analogiška procedūra kartota ir išoriniam sluoksniui su 3 neuronais, kur įeitys — praeitame žingsnyje gauti rezultatai po aktyvacijos funkcijos.

Lentėje 5 matyti WEKA sitemos ir skaičiuoklės klasifikavimo tikimybės daug kur sutampa. Nesutapimai paryškinti. Jos sutapti visos negali, kadangi skaičiuoklė palaiko ribotą kiekį skaičių po kablelio, dėl ko atsitinka apvalinimo netikslimų.

3 lentelė. Paslėptojo sluoksnio svoriai. Svoriai lentelėje pateikti 2 skaičių po kablelio tikslumu

	Neuronas 1	Neuronas 2	Neuronas 3	Neuronas 4	Neuronas 5
poslinkis	2,12	-6,10	-5,68	1,81	-1,95
sepallength	0,59	-1,37	-1,22	0,83	-0,37
sepalwidth	-1,76	-2,53	-2,36	-1,33	2,15
petallength	2,59	7,82	7,30	2,03	-2,84
petalwidth	2,83	7,78	7,24	1,68	-3,09

4 lentelė. Išorinio sluoksnio svoriai. Svoriai lentelėje pateikti 2 skaičių po kablelio tikslumu

	Išorinis neuronas 1	Išorinis neuronas 2	Išorinis neuronas 3
poslinkis	0,52	-0,09	-4,83
Neuronas 5	4,93	-5,40	-4,93
Neuronas 4	-2,11	2,12	-0,87
Neuronas 3	-1,70	-6,45	6,01
Neuronas 2	-1,79	-7,03	6,47
Neuronas 1	-4,25	3,54	0,76

5 lentelė. WEKA ir skaičiuoklės klasifikavimo prognozių tikimybės. Lentelėje tikimybės pateiktos 2 skaičių po kablelio tikslumu

Iris-setosa		**Iris-versicolor**		**Iris-virginica**	
WEKA prog,	**Excel prog,**	**WEKA prog,**	**Excel prog,**	**WEKA prog,**	**Excel prog,**
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,00	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,99	0,99	0,01	0,01	0,00	0,00
0,00	0,00	0,99	0,99	0,01	0,01
0,00	0,00	0,99	0,99	0,01	0,01
0,00	0,00	**0,98**	**0,99**	**0,02**	**0,01**
0,00	0,00	0,99	0,99	0,01	0,01
0,00	0,00	0,98	0,98	0,02	0,02
0,00	0,00	0,99	0,99	0,01	0,01
0,00	0,00	**0,98**	**0,99**	**0,02**	**0,01**
0,02	**0,01**	0,98	0,98	0,00	0,00
0,00	0,00	0,99	0,99	0,01	0,01
0,01	0,01	0,99	0,99	0,01	0,01
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	**0,08**	**0,02**	**0,92**	**0,98**
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	0,00	0,00	1,00	1,00
0,00	0,00	0,00	0,00	1,00	1,00

3. Išvados

Irisų duomenų rinkinys mažas ir paprastas, todėl nemažai dalis paremtrų, kaip paslėptų neuronų skaičius, inercija, neturi didelės įtakos neuroninio tinklo mokymui, tačiau mokymosi greičio parametras išlieka reikšmingas. Geriausias modelio tikslumas gautas trečioje užduočių sekoje, kur 120 įrašų naudojami mokymui ir likę 30 — testavimui. Skaičiuoklėje galima atkurti modelį, galintį prognozuoti tikimybinės klasifikavimo reikšmes su ganėtinai maža paklaida lyginant su WEKA sistemos tais pačiais modelio parametrais.