VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Tiesioginio sklidimo DNT naudojant sistemą WEKA

3 užduotis

Atliko: 4 kurso 1 grupės studentė
Rosita Raišuotytė

Lapkričio 18, 2023

TURINYS

1. Įvadas

1.1. Tikslas

Užduoties tikslas - išmokyti neuroninį tinklą teisingai klasifikuoti duomenis naudojant sistemą WEKA, bei sukurti neuroninį tinklą Microsoft Excel aplinkoje naudojant WEKA sistemoje gautus svorius.

1.2. Uždavinai

- Paruošti duomenis irisų duomenis išskiriant į mokymosi-testavimo ir naujus duomenų aibes.
- Sukonstruoti WEKA sistemoje tris sekas:
  - daugiasluoksnio perceptrono apmokymui;
  - apmokyto daugiasluoksio perceptrono naujų duomenų klasifikavimui;
  - daugiasluoksnio perceptono duomenims klasifikuoti ir testuoti.
- Rasti geriausius daugiasluoksio perceptrono parametrus atliekant bandymus su pirmąja seka.
- Suklasifikuoti naujus duomenis naudojantis apmokytu neuroniniu tinklu.
- Apmokyti ir ištestuoti daugiasluoksnį perceproną, bei gautus svorius panaudoti skaičiuojant neuronų išėjimus Microsoft Excel skaičiuoklėje.
- Palyginti gautus rezultatus WEKA sistemoje su paskaičiuotais Microsoft Exel programoje.

2. Duomenys

Naudojama irisų duomenų aibė turinti 150 įrašų, po 50 kiekvienai klasei.

2.1. Duomenų paruošimas

Duomenys yra perskirti į mokymosi-testavimo ir naujų duomenų aibes. Mokymosi-testavimo duomenų aibėje yra 120 įrašų, po 40 įrašų kiekvienos klasės. Naujų duomenų aibę sudaro 30 įrašų, po 10 kiekvienos klasės.

2.2. Duomenų požymiai

Pagal užduoties 0 (nulinį) variantą, pasirinkti požymiai:

- sepallength
- sepalwidth
- petallength

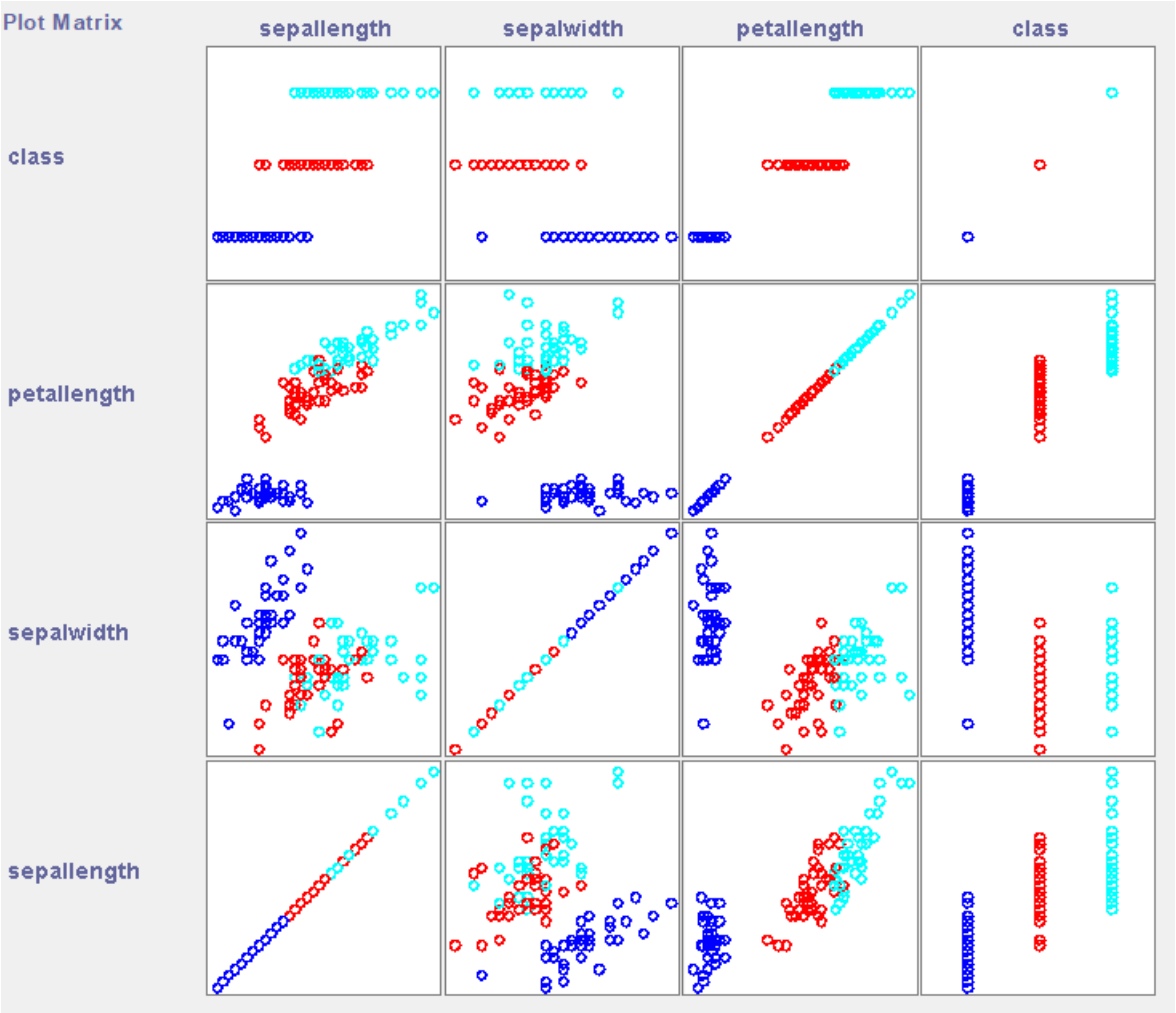
2.3. Duomenų požymiai Dekarto koordinačių sistemoje

Dekarti koordinačių sistemoje duomenys pažymėti:

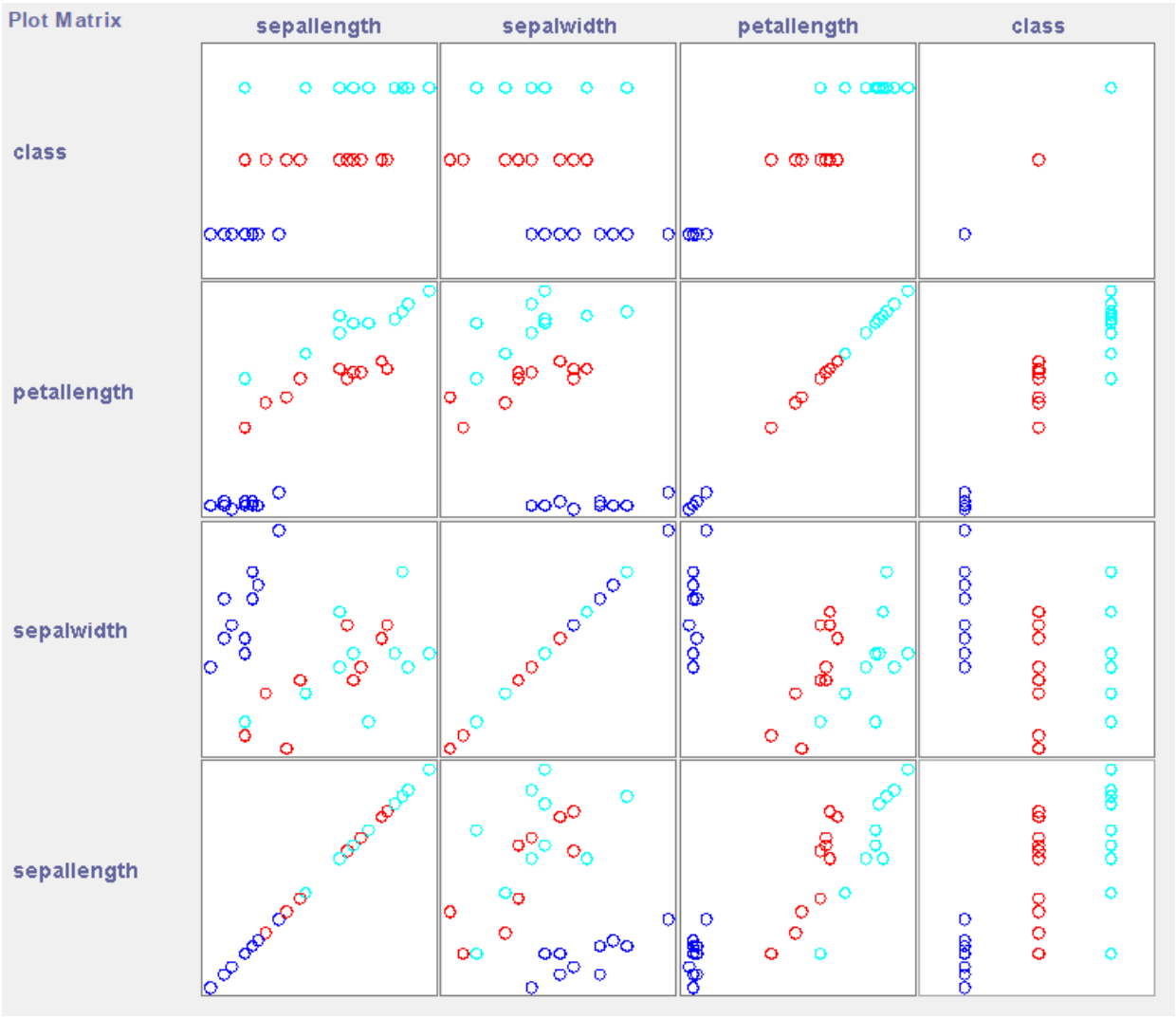
- mėlyna spalva - Iris-setosa klasė;
- raudona - Iris-versicolor;

- žydra - Iris-virginica.

Iš 1 pav. ir matome, 2 pav. kad Iris-setosa klasės požymiai yra išsiskiria nuo kitų klasių, vadinasi, šią klasę modeliui bus lengviau išmokti ir atpažinti. Likusių dviejų klasių požymiai persipynę, juos modeliui bus sunkiau atpažinti. 2 pav. naujų duomenų yra mažiau ir požymiai mažiau persipynę, todėl modeliui klasifikuoti juos bus lengviau nei mokymosi-testavimo duomenis.



1 pav. Mokymosi-testavimo duomenų aibės požymių porų vaizdai Dekarto koordinačių sistemoje



2 pav. Naujų duomenų aibės požymių porų vaizdai Dekarto koordinačių sistemoje

3. Užduočių sekos

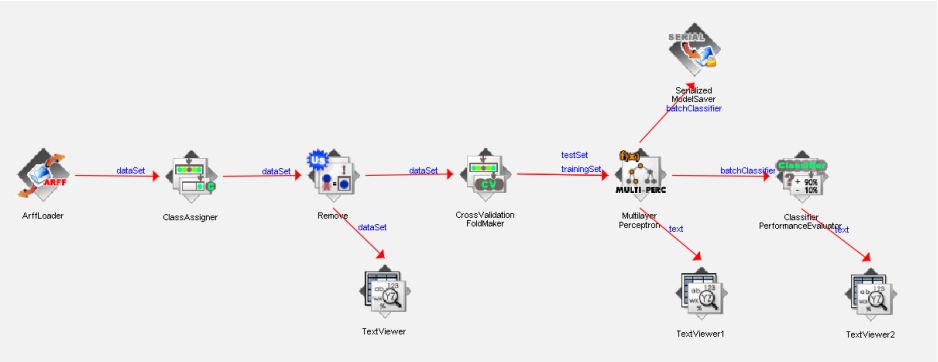
3.1. Pirmoji seka – daugiasluoksnio perceptrono apmokymas

3 pav. matoma užduočių seka skirta apmokyti daugiasluoksnį perceptroną keičiant jo parametrus ir išsaugoti modelį su kuriuo bandymo rezultatai buvo geriausi. Užduočių sekoje matomi elementai:

- ArffLoader – skaito arff tipo įvesties duomenų failą, šiuo atveju iris_train_test.arff.
- ClassAssigner – reikalingas nurodyti, kuris duomenų požymis yra klasė.
- Remove – panaikina nereikalingus požymius, pagal turimą užduoties variantą tai ketvirtas (petalwidth) požymis.
- CrossValidationFoldMaker – sukuria skirtingus kryžminės patikros duomenų blokus bei dalina duomenis į mokymosi ir testavimo aibes santykiu 80:20. Nustatyta reikšmė 5.
- MultilayerPerceptor – daugiasluoksnis perceptronas, kuriame galima keisti įvairius parametrus. Paketo dydį (angl. batchSize) nustatome 10.
- SerializedModelSaver – skirtas išsaugoti apmokytą daugiasluoksnį perceptroną. Kadangi

CrossValidationFoldMaker reikšmė nustatyta 5, išsaugomi 5 modeliai apmokyti su skirtingai padalintais mokymo testavimo duomenimis.

- ClassifierPerformanceEvaluator – įvertina klasifikavimo rezultatus.
- Textviewer – gauna ir atvaizduoja tekstinius duomenis.

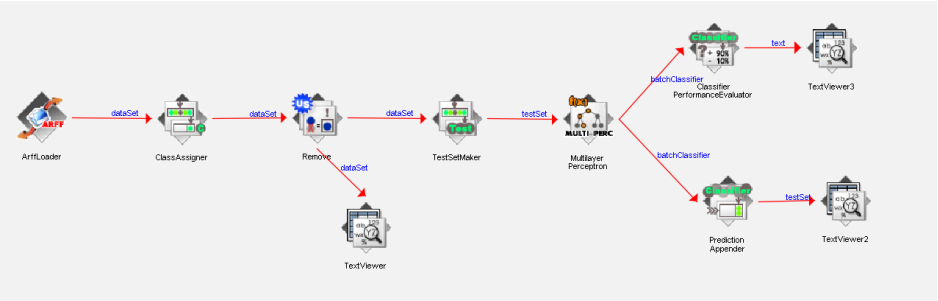


3 pav. Daugiasluoksnio peceptrono apmokymo užduočių seka WEKA sistemoje

3.2. Antroji seka – naujų duomenų klasifikavimas su apmokytu DNT

4 pav. matoma užduočių seka skirta naujiems duomenims klasifikuoti su pirmoje sekoje apmokytu ir išsaugotu modeliu. Čia naudojamas duomenų failas iris_new.arff. Pirmoje užduočių sekoje nebuvę elementai:

- TestSetMaker – iš pateiktų duomenų sukuria testavimo duomenų aibę.
- PredictionAppender – pateiktiems duomenims priskiria modelio prognozuojamas klases arba klasių tikimybes.

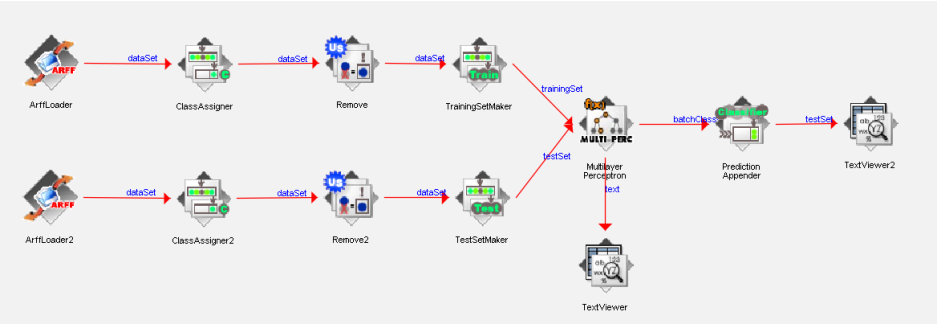


4 pav. Naujų duomenų klasifikavimo su apmokytu DNT užduočių seka WEKA sistemoje

3.3. Trečioji seka – DNT apmokymas ir testavimas

5 pav. matoma užduočių seka, kurioje daugiasluoksnis perceptronas su vienais duomenimis yra apmokomas ir su kitais duomenimis testuojamas. Užduočių sekos elementai:

- ArffLoader – skaito iris_train_test.arff duomenų failą.
- ArffLoader2 – skaito iris_new.arff duomenų failą.
- TrainingSetMaker – iš pateiktų duomenų sukuria mokymosi duomenų aibę.

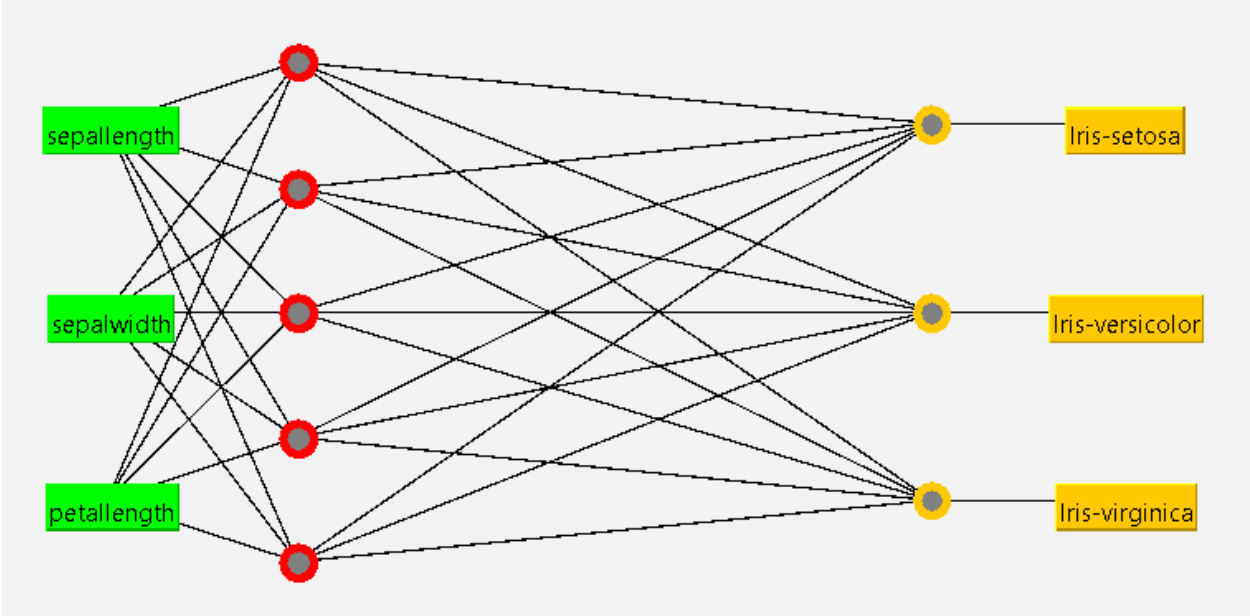


5 pav. Daugiasluoksnio peceptrono apmokymo ir testavimo užduočių seka WEKA sistemoje

4. Rezultatai

4.1. Parametrų radimas pirmąjai sekai

Buvo atlikta 18 suplanuotų bandymų keičiant paslėpto sluoksnio neuronų skaičių, mokymosi greitį bei mokymosi pagreitį (momentum). Kiti parametrai buvo fiksuoti: epochų skaičius - 500, paketo dydis - 10. Bandymai buvo atlikti taikant visas galimas kombinacijas, kai paslėpto sluoksnio neuronų skaičius imamams iš aibės {3; 5}, mokymosi greitis - {0,2; 0,5; 0,9}, o pagreitis - {0,01; 0,1; 0,5}. Šių bandymų rezultatai matomi 1 lentelėje. Taip pat buvo atlikti ir atsitiktiniai bandymai, kurių rezultatai nebus pateikti, nes nei vienas bandymas neviršijo 0,95 klasifikavimo tikslumo. Geriausi rezultatai gauti su šiais parametrais: paslėpto sluoksnio neuronų skaičius - 5, mokymosi greitis - 0,5, mokymosi pagreitis - 0,1. Gautas klasifikavimo tikslumas 0,95, kaip ir daugymoje bandymų, todėl buvo atsižvengta ir į santykinę paklaidą, kurios reikšmė esant tokiam tikslumui mažiausia buvo 0,0507. Dirbtinis neuroninis tinklas su 5 neuronais paslėptame sluoksnyje pavaizduotas 6 pav.



6 pav. Dirbtinis neuroninis tinklas su 5 neuronais paslėptame sluoksnyje

1 lentelė. Klasifikavimo tikslumas ir santykinė paklaida su skirtingais parametrais

Nr.	Neuronų skaičius paslėptame sluoksnyje	Mokymosi gretis	Mokymosi pagretis (momen-tum)	Klasifikavio tikslumas	Santykinė paklaida (angl. mean error)
1	3	0,2	0,01	0,950	0,0582
2	3	0,2	0,1	0,950	0,0573
3	3	0,2	0,5	0,950	0,0539
4	3	0,5	0,01	0,950	0,0529
5	3	0,5	0,1	0,950	0,0527
6	3	0,5	0,5	0,933	0,0573
7	3	0,9	0,01	0,942	0,0550
8	3	0,9	0,1	0,942	0,0564
9	3	0,9	0,5	0,925	0,0610
10	5	0,2	0,01	0,950	0,0570
11	5	0,2	0,1	0,950	0,0562
12	5	0,2	0,5	0,950	0,0524
13	5	0,5	0,01	0,950	0,0511
14	**5**	**0,5**	**0,1**	**0,950**	**0,0507**
15	5	0,5	0,5	0,933	0,0583
16	5	0,9	0,01	0,950	0,0537
17	5	0,9	0,1	0,942	0,0543
18	5	0,9	0,5	0,925	0,0551

4.2. Naujų duomenų klasifikavimas pagal antrąja seką

Nauji duomenys buvo klasifikuoti su jau apmokytu modeliu, kurio parametrai: paslėpto sluoksnio neuronų skaičius - 5, mokymosi greitis - 0,5, mokymosi pagreitis - 0,1, epochų skaičius - 500, paketo dydis - 10. 2 lentelėje pateikti naujų duomenų klasifikavimo rezultatai. Klasifikavimo tikslumas lygus 1,0. Nors mokymo metu modelio klasifikvimo tikslumas buvo 0,95, naujus duomenis jis suklasifikavo nesuklysdamas.

2 lentelė. Naujų duomenų įrašai su modelio gautais rezultatais

Duomenų įrašas su klase				Spėjama klasė	Spėjamos tikimybės kiekvienai klasei		
5,1	3,5	1,4	Iris-setosa	Iris-setosa	0,990745	0,009252	0,000003
4,9	3,0	1,4	Iris-setosa	Iris-setosa	0,985564	0,014433	0,000003
4,7	3,2	1,3	Iris-setosa	Iris-setosa	0,990197	0,009801	0,000003
4,6	3,1	1,5	Iris-setosa	Iris-setosa	0,987997	0,012	0,000003
5,0	3,6	1,4	Iris-setosa	Iris-setosa	0,991543	0,008455	0,000003
5,4	3,9	1,7	Iris-setosa	Iris-setosa	0,991108	0,008889	0,000003
4,6	3,4	1,4	Iris-setosa	Iris-setosa	0,991277	0,00872	0,000003
5,0	3,4	1,5	Iris-setosa	Iris-setosa	0,989641	0,010357	0,000003
4,4	2,9	1,4	Iris-setosa	Iris-setosa	0,987317	0,012679	0,000003
4,9	3,1	1,5	Iris-setosa	Iris-setosa	0,985992	0,014005	0,000003
7,0	3,2	4,7	Iris-versicolor	Iris-versicolor	0,012725	0,984099	0,003176
6,4	3,2	4,5	Iris-versicolor	Iris-versicolor	0,01362	0,984307	0,002073
6,9	3,1	4,9	Iris-versicolor	Iris-versicolor	0,005559	0,938235	0,056206
5,5	2,3	4,0	Iris-versicolor	Iris-versicolor	0,008055	0,990388	0,001558
6,5	2,8	4,6	Iris-versicolor	Iris-versicolor	0,006769	0,980844	0,012387
5,7	2,8	4,5	Iris-versicolor	Iris-versicolor	0,004113	0,945636	0,050251
6,3	3,3	4,7	Iris-versicolor	Iris-versicolor	0,009184	0,977972	0,012844
4,9	2,4	3,3	Iris-versicolor	Iris-versicolor	0,021321	0,978481	0,000198
6,6	2,9	4,6	Iris-versicolor	Iris-versicolor	0,008566	0,984764	0,00667
5,2	2,7	3,9	Iris-versicolor	Iris-versicolor	0,014041	0,985299	0,00066
6,3	3,3	6,0	Iris-virginica	Iris-virginica	0,000206	0,001615	0,998179
5,8	2,7	5,1	Iris-virginica	Iris-virginica	0,000251	0,007021	0,992728
7,1	3,0	5,9	Iris-virginica	Iris-virginica	0,000206	0,002036	0,997758
6,3	2,9	5,6	Iris-virginica	Iris-virginica	0,000202	0,002309	0,99749
6,5	3,0	5,8	Iris-virginica	Iris-virginica	0,0002	0,001921	0,997878
7,6	3,0	6,6	Iris-virginica	Iris-virginica	0,000202	0,001495	0,998304
4,9	2,5	4,5	Iris-virginica	Iris-virginica	0,000594	0,12478	0,874625
7,3	2,9	6,3	Iris-virginica	Iris-virginica	0,000198	0,00158	0,998222
6,7	2,5	5,8	Iris-virginica	Iris-virginica	0,000183	0,001908	0,997909
7,2	3,6	6,1	Iris-virginica	Iris-virginica	0,000227	0,001895	0,997878

4.3. Gauti DNT svoriai vykdant trečiąja seką

Sukurtas daugiasluoksnis perceptronas su parametrais: paslėpto sluoksnio neuronų skaičius - 5, mokymosi greitis - 0,5, mokymosi pagreitis - 0,1, epochų skaičius - 500. Daugiasluoksnis perceptronas buvo apmokytas ir gauti šie svorių rinkiniai: paslėpto sluooksnio neuronų svoriai 3 lentelėje, išeities sluoksnio neuronų svoriai 4 lentelėje. Svoriai dėl patogumo buvo apvalinti 10^{-6} tikslumu.

3 lentelė. Paslėpto sluoksnio neuronų svoriai

	threshold	sepallength	sepalwidth	petallength
node 3	-3,919716	-1,413302	-2,611756	14,746836
node 4	3,439212	2,011974	1,922517	-14,211757
node 5	2,764270	1,523271	-2,293532	5,494378
node 6	-3,132513	-1,907848	2,427720	-6,323071
node 7	-2,520762	-1,750081	-0,031374	8,720548

4 lentelė. Išeities sluoksnio neuronų svoriai

	threshold	node 3	node 4	node 5	node 6	node 7
node 0	-0,852325	-2,202412	1,296916	-5,412643	4,815116	-2,904716
node 1	-1,129573	-7,159035	4,942018	6,103361	-8,922185	-3,219589
node 2	-3,288432	4,945434	-7,890764	0,597928	-3,690957	3,597661

4.4. Microsoft Excel aplinkoje gautas neuroninis tinklas

Microsoft Excel aplinkoje atlikti žingsniai su kuriais gautas neuroninis tinklas:

- Įsikelti naujų duomenų įrašai iš iris_new.arff duomenų failo;
- Įsikeltos modelio iš trečios užduočių sekos gautos klasių tikimybės ir spėjama klasė;
- Normalizuojami įėjimo duomenys, kad priklausytų intervale [-1,1];
- Sukuriamos ir užpildomos svorių lentelės su gautais svoriais WEKA sistemoje iš trečiosios užduočių sekos;
- Suskaičiuoti paslėpto sluoksnio neuronų išėjimų sumos, o vėliau pritaikyta sigmoidinė aktyvacijos funkcija
- Suskaičiuoti išėjimo neuronų išėjimų sumos, pritaikyta sigmoidinė aktyvacijos funkcija (gautos reikšmės yra klasių tikimybės).

Visus skaičiavimus ir rezultatus galite matyti paspaudę nuorodą čia. Gautus rezultatus matote 5 lentelėje, kur 0 atitiktų Iris-setosa klasę, 1 - Iris-versicolor, 2 - Iris-virginica. Lentelėje pateikti rezultatai yra suapvalinti dėl mažesnio kiekio skaičių. Rezultatai gavosi labai panašūs, MS Excel skaičiuoklėje yra pateiktos paklaidos kiekvienam įrašui ir paklaidos vidurkis. Matoma, kad Iris-setosa klasės tikimybių vidurkinė paklaida yra apie -0,0011, Iris-versicolor apie 0,015 ir Iris-virginica apie -0,0134. Paklaidos tikriausiai atsirado dėl netikslių ilgų skaičių po kablelio.

5 lentelė. WEKA sistemoje ir MS Excel skaičiuoklėje gautos tikimybės

	Tikimybės gautos WEKA			Tikimybės gautos MS Excel		
Nr.	0	1	2	0	1	2
1	0,99147	0,00853	0,00000	0,99438	0,00671	0,00000
2	0,98726	0,01274	0,00000	0,99313	0,00882	0,00000
3	0,99143	0,00857	0,00000	0,99425	0,00691	0,00000
4	0,98932	0,01069	0,00000	0,99371	0,00780	0,00000
5	0,99218	0,00782	0,00000	0,99453	0,00647	0,00000
6	0,99117	0,00883	0,00000	0,99447	0,00658	0,00000
7	0,99215	0,00785	0,00000	0,99447	0,00656	0,00000
8	0,99039	0,00961	0,00000	0,99415	0,00709	0,00000
9	0,98928	0,01072	0,00000	0,99350	0,00815	0,00000
10	0,98723	0,01277	0,00000	0,99327	0,00859	0,00000
11	0,00131	0,99597	0,00272	0,00136	0,99657	0,00224
12	0,00191	0,99707	0,00103	0,00202	0,99859	0,00082
13	0,00030	0,81451	0,18520	0,00024	0,70912	0,26692
14	0,00231	0,99679	0,00090	0,00143	0,99366	0,00382
15	0,00064	0,97148	0,02788	0,00046	0,90979	0,07025
16	0,00042	0,90211	0,09748	0,00027	0,73332	0,25158
17	0,00066	0,97974	0,01960	0,00067	0,97936	0,01654
18	0,01793	0,98205	0,00002	0,01402	0,99976	0,00002
19	0,00087	0,98831	0,01083	0,00070	0,97248	0,01990
20	0,00522	0,99462	0,00015	0,00460	0,99938	0,00019
21	0,00001	0,00507	0,99492	0,00001	0,00480	0,99686
22	0,00003	0,01459	0,98538	0,00002	0,00984	0,99295
23	0,00002	0,00635	0,99364	0,00001	0,00561	0,99629
24	0,00002	0,00675	0,99324	0,00001	0,00581	0,99613
25	0,00001	0,00587	0,99411	0,00001	0,00529	0,99652
26	0,00001	0,00478	0,99520	0,00001	0,00465	0,99701
27	0,00009	0,12856	0,87135	0,00005	0,04357	0,96181
28	0,00001	0,00506	0,99493	0,00001	0,00481	0,99689
29	0,00002	0,00599	0,99400	0,00001	0,00539	0,99647
30	0,00001	0,00578	0,99421	0,00001	0,00531	0,99651

5. Išvados

- Irisų duomenys geriausiai suklasifikuoti buvo sudarius neuroninį tinklą iš vieno paslėpto sluoksnio su 5 neuronais, kai mokymosi greitis buvo lygus 0,5, o pagreitis (momentum) – 0,1. Gautas mokymosi tikslumas 0,95, o testavimo su naujais duomenimis – 1,0.

- WEKA sistemos neuroninį tinklą galima įgyvendinti ir MS Excel aplinkoje, tačiau gaunami rezultatai gali turėti nedidelę paklaidą, dėl skaičių po kablelio.