VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

STUDIJŲ PROGRAMA DUOMENŲ MOKSLAS 4 KURSAS

Tiesioginio sklidimo DNT naudojant sistemą WEKA

Darbo aprašas

Autorius: Austėja Ona Plančiūnaitė
VU el. p.: ona.plančiūnaitė@mif.stud.vu.lt

Vilnius
2024

Turinys

1 Įvadas

Praktinės užduoties tikslas yra išmokyti neuroninį tinklą klasifikuoti duomenis naudojant sistemą WEKA. Darbe sukonstruojamos trys skirtingos WEKA programos užduočių sekos, atliekamas neuroninio tinklo parametrų parinkimo tyrimas. Be to, neuroninio tinklo išėjimo reikšmės perskaičiuojamos MS Excel programoje, palyginami WEKA ir Excel gauti rezultatai.

2 Praktinė užduotis

2.1 Tikslas

Praktinės užduoties tikslas yra išmokyti neuroninį tinklą teisingai klasifikuoti duomenis naudojant sistemą WEKA.

2.2 Uždaviniai

- Duomenų paruošimas, paskirstymas į mokymo ir testavimo aibes.
- Užduočių sekų sukonstravimas WEKA sistemoje.
- Neuroninio tinklo parametrų parinkimas. Reikia parinkti tokius paslėptų neuronų skaičius (hiddenLayers), mokymo greičio parametro (learningRate) bei momentum reikšmes, kad tinklas geriausiai išmoktų klasifikuoti duomenis.
- Antros užduočių sekos konstravimas naujiems duomenims klasifikuoti.
- Neuronų išėjimo reikšmių perskaičiavimas MS Excel programoje.
- Palyginti gautus rezultatus WEKA ir MS Excel programose.

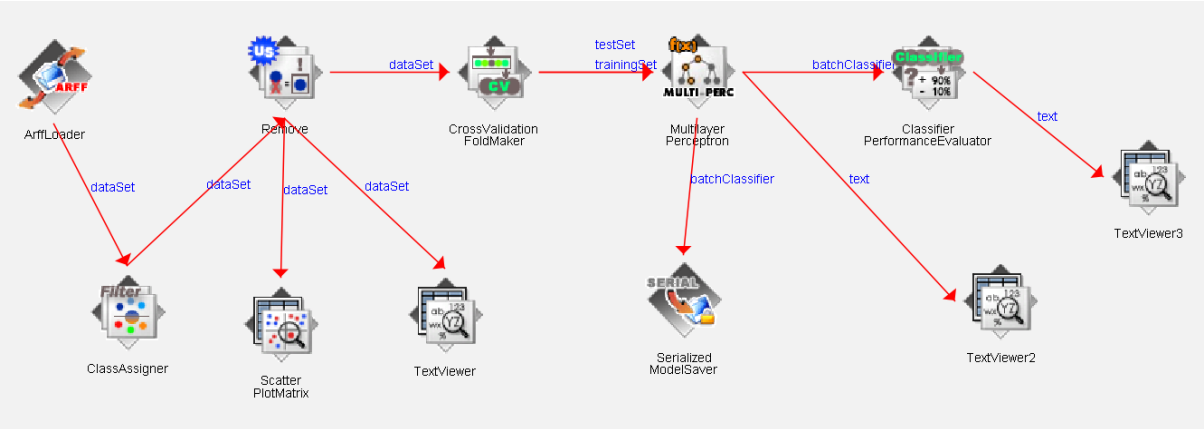
2.3 Duomenys

Praktiniam darbui atlikti naudojamas „Irisų“ duomenų rinkinys. Šiame duomenų rinkyje išviso yra 150 įrašų, 4 požymiai ir 3 galimos klasės reikšmės: Setosa, Versicolor ir Virginica. Tačiau darbe naudojami duomenys klasifikuojami tik pagal 3 požymius: sepallength, sepalwidth ir petallength. Be to, atlikdami tyrimą šią duomenų aibę padalinsime į mokymo ir testavimo dalis santykiu 80:20. Vadinasi, 120 įrašų priklauso mokymo aibei ir 30 testavimo. Mokymo duomenų rinkinyje yra po 40 kiekvienos klasės duomenų, o testavimo atitinkamai po 10 įrašų.

3 Uždavinių sekos WEKA sistemoje

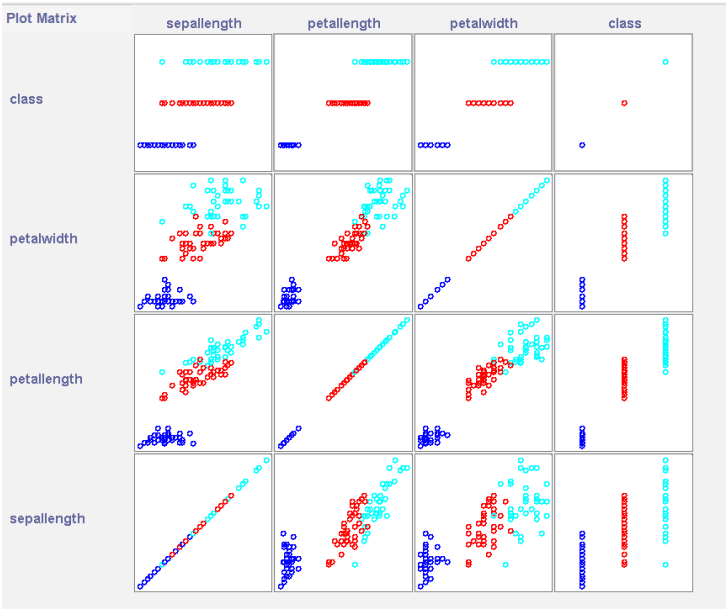
3.1 Pirmoji seka

Sukonstruota užduočių seka daugiasluoksniam perceptronui apmokyti vaizduojama 1 paveikslėlyje. „ArffLoader“ komponentėje nurodomas mokymo duomenų failas, nereikalingą požymį išmetame naudodami „Remove“ komponentę. Kryžminė patikra atliekama „CrossValidation“ komponentėje, šioje komponentėje pakeičiame kryžminės patikros blokų skaičių į 5. Naudodami „Serialized ModelSaver“ išsaugome apmokytą modelį. „MultiLayer Perceptron“ yra daugiasluoksnio perceptrono komponentė, skirta klasifikavimo uždaviniams spręsti, joje pakeičiame paketo dydį (batchSize) į 10. Klasifikavimo tikslumą galima pamatyti naudojant „Classifier Performance Evaluator“. Norint tinkamai parinkti parametrus neuroniniam tinklui, kad jis geriausiai išmoktų klasifikuoti duomenis, „MultiLayer Perceptron“ komponentėje keičiame parametrus ir tikriname jų įtaką klasifikavimo tikslumui. Atliekame tyrimą su skirtingais paslėptų neuronų skaičiais, mokymo greičio ir momentum reikšmėmis. Rezultatai pateikiami 1 ir 2 lentelėse, kur klasifikavimo tikslumą vertiname pagal teisingai suklasifikuotų duomenų įrašų skaičių.



1 pav.: Pirmoji užduočių seka.

Taip pat, pirma pažiūrime duomenų požymių porų pasiskirstymas Dekarto koordinačių sistemoje naudojant „Scatter PlotMatrix“ komponentę. Gauta matrica, vaizduojanti šių porų išsidėstymą, pateikiama 2 paveiksle. Čia šviesiai mėlyna spalva pažymėta „virginica“ klasė, o raudona ir mėlyna atitinkamai žymi „setosa“ ir „versicolor“ klases.



2 pav.: Duomenų požymių porų pasiskirstymas Dekarto sistemoje naudojant mokymo duomenis.

Remiantis 1 lentelėje pateiktais rezultatais, galime teigti, jog neuroniniam tinklui su vienu paslėptu neuronų sluoksniu geriausią klasifikavimo tikslumą galima pasiekti naudojant kelias skirtingas modelio parametrų variacijas. Pavyzdžiui, naudojant tik 2 neuronus paslėptame sluoksnyje kai mokymosi greitis 0,3 ir momentum reikšmė 0,2, klasifikavimo modelis pasiekia 95 % tikslumą, teisingai suklasifikuodamas 114 iš 120 duomenų įrašų. Analogiškai, pasirinkus 3 neuronus pirmame sluoksnyje, su visomis mokymosi greičio ir momentum reikšmėmis modelis irgi įgyja 95 %.

Neuronų skaičius pirmame sluoksnyje	Mokymosi greitis	Momentum	Tikslumas
2	0.3	0.2	95%
2	0.6	0.2	94.17%
2	0.9	0.2	94.17%
2	0.3	0.4	94.17%
2	0.6	0.4	94.17%
2	0.9	0.4	95%
3	0.3	0.2	95%
3	0.6	0.2	95%
3	0.9	0.2	95%
3	0.3	0.4	95%
3	0.6	0.4	95%
3	0.9	0.4	95%
4	0.3	0.2	95%
4	0.6	0.2	95%
4	0.9	0.2	94.17%
4	0.3	0.4	95%
4	0.6	0.4	95%
4	0.9	0.4	95%

1 lentelė: Klasifikiavimo tikslumas, kai yra vienas paslėptas neuronų sluoksnis.

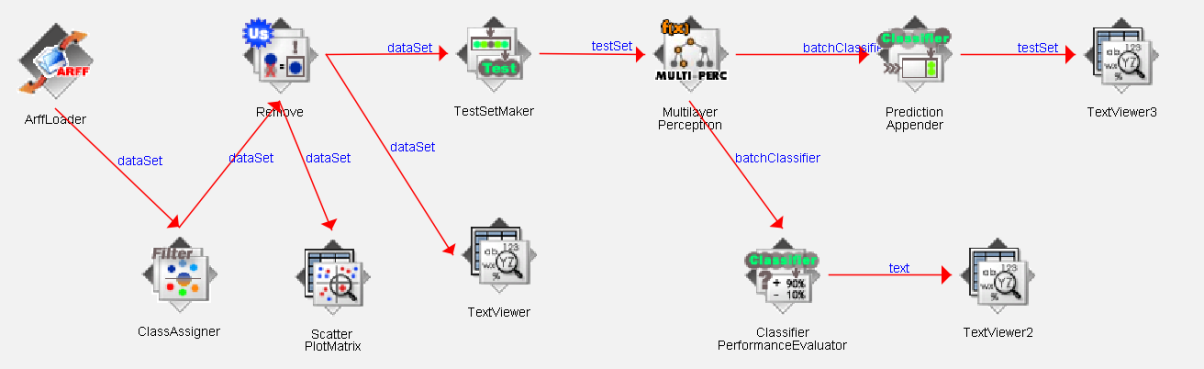
Remiantis 2 lentelėje pateiktais rezultatais, galime teigti, kad neuroniniam tinklui su dviem paslėptais sluoksniais, geriausią klasifikavimo tikslumą pasiekti galima pasirinkus įvairias modelio parametrų variacijas. Pavyzdžiui, parinkus 4 neuronus pirmame sluoksnyje, antrame sluoksnyje 8, mokymosi greitį 0,6 ir momentum reikšmę 0,2, klasifikavimo modelis pasiekia 95,83 % tikslumą, teisingai suklasifikuodamas 115 iš 120 duomenų įrašų.

Neuronų skaičius pirmame sluoksnyje	Neuronų skaičius antrame sluoksnyje	Mokymosi greitis	Momentum	Tikslumas
2	4	0,3	0,2	94,17 %
2	4	0,6	0,2	95 %
2	4	0,9	0,2	95 %
2	4	0,3	0,4	95 %
2	4	0,6	0,4	95 %
2	4	0,9	0,4	95 %
3	6	0,3	0,2	95 %
3	6	0,6	0,2	95 %
3	6	0,9	0,2	95 %
3	6	0,3	0,4	94,17 %
3	6	0,6	0,4	95 %
3	6	0,9	0,4	95 %
4	8	0,3	0,2	95 %
4	8	0,6	0,2	95,83 %
4	8	0,9	0,2	95,83 %
4	8	0,3	0,4	95 %
4	8	0,6	0,4	95,83 %
4	8	0,9	0,4	95,83 %

2 lentelė: Neuronų tinklo su 2 paslėptais sluoksniais parametrų lentelė ir tikslumas

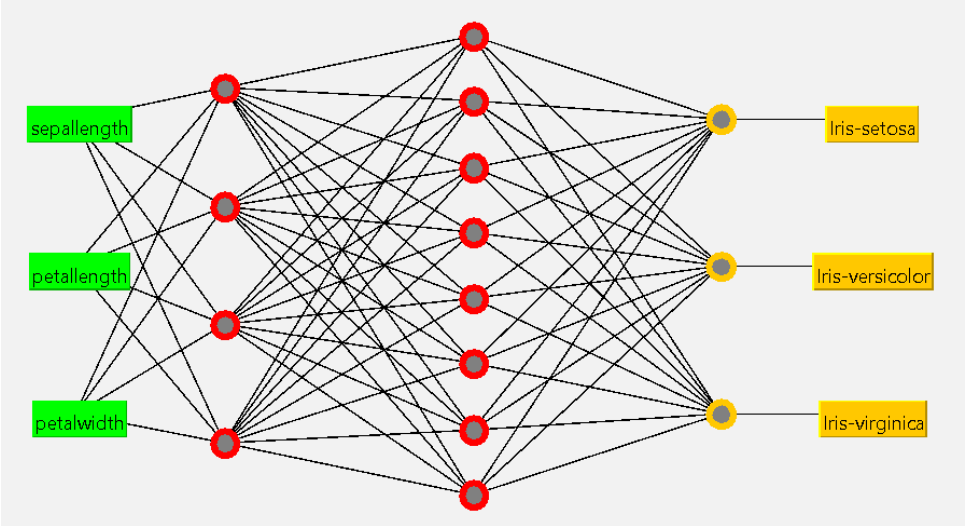
3.2 Antroji seka

Antroje užduočių sekoje naudojamas 3 paveiksle pavaizduotas neuroninis tinklas. Turime 2 paslėptus neuronų sluoksnius: vienas su 4-iais neuronais, o kitas su 8-iais. Mokymosi greitis nustatytas 0,6, o momentum reikšmė 0,2. Antroji sukonstruota užduočių seka naudoja anksčiau sukurtą ir išsaugotą tinklo modelį. Čia naudojami testavimo duomenys ir jiems priskiriama klasė. Naudodami „TestSetMaker“ komponentę duomenis priskiriame testavimui. Pridedant „Prediction Appender“ galime matyti su kokia tikimybe įrašas buvo priskirtas kiekvienai klasei.



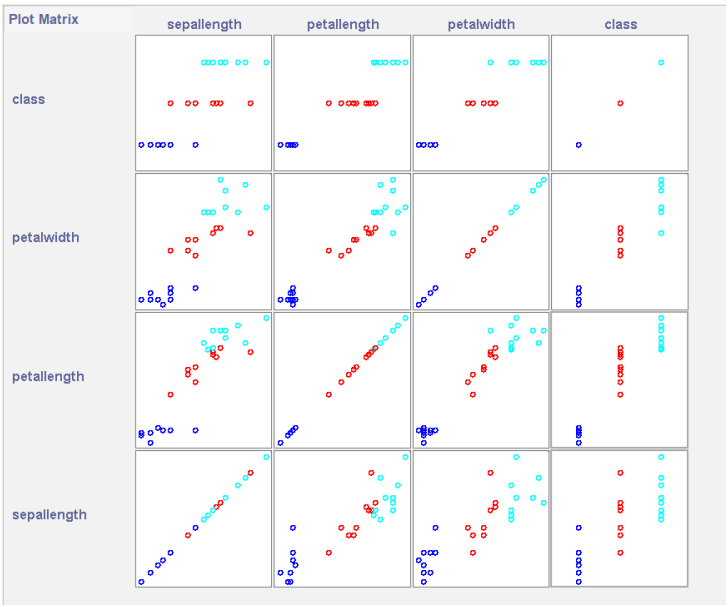
3 pav.: Antroji užduočių seka

Antroje užduočių sekoje naudojamas 4 paveiksle pavaizduotas neuroninis tinklas. Turime 2 paslėptus neuronų sluoksnius: viename yra 4 neuronai, o kitame 8.



4 pav.: Neuroninio tinklo struktūra.

Taip pat, vėl pažiūrime duomenų požymių porų pasiskirstymas Dekarto koordinačių sistemoje naudojant „Scatter PlotMatrix“ komponentę. Gauta matrica, vaizduojanti šių porų išsidėstymą, pateikiama 2 paveiksle. Čia šviesiai mėlyna spalva pažymėta „virginica“ klasė, o raudona ir mėlyna atitinkamai žymi „setosa“ ir „versicolor“ klases.



5 pav.: Duomenų požymių porų pasiskirstymas Dekarto sistemoje naudojant testavimo duomenis.

Klasifikavimo modelio rezultatai pateikiami 6 paveiksle. Kiekvienam duomenų įrašui pateikiamos klasių priskyrimo tikimybių reikšmės.

```
@attribute sepallength numeric
@attribute petallength numeric
@attribute petalwidth numeric
@attribute class {Iris-setosa,Iris-versicolor,Iris-virginica}
@attribute MultilayerPerceptron_prob_Iris-setosa numeric
@attribute MultilayerPerceptron_prob_Iris-versicolor numeric
@attribute MultilayerPerceptron_prob_Iris-virginica numeric

@data
6.4,5.6,2.2,Iris-virginica,0,0.033643,0.966357
5.9,5.1,1.8,Iris-virginica,0,0.046262,0.953738
6.1,4.9,1.8,Iris-virginica,0,0.068424,0.931576
6,4.8,1.8,Iris-virginica,0,0.081655,0.918345
6.7,5.8,1.8,Iris-virginica,0,0.036434,0.963566
6.1,5.6,1.4,Iris-virginica,0,0.049428,0.950572
7.4,6.1,1.9,Iris-virginica,0,0.035276,0.964724
6.3,5.6,2.4,Iris-virginica,0,0.032766,0.967234
6.9,5.1,2.3,Iris-virginica,0,0.038487,0.961513
6.4,5.3,1.9,Iris-virginica,0,0.040794,0.959206
4.9,1.5,0.1,Iris-setosa,0.990936,0.008856,0.000208
5.1,1.5,0.2,Iris-setosa,0.990932,0.00886,0.000208
4.4,1.4,0.2,Iris-setosa,0.990061,0.00973,0.000209
5.1,1.5,0.3,Iris-setosa,0.990667,0.009125,0.000208
4.6,1,0.2,Iris-setosa,0.991219,0.008574,0.000207
5.7,1.5,0.4,Iris-setosa,0.991133,0.00866,0.000207
5.1,1.5,0.4,Iris-setosa,0.990332,0.00946,0.000209
4.4,1.3,0.2,Iris-setosa,0.990374,0.009418,0.000208
```

6 pav.: Klasifikavimo rezultatai kiekvienam duomenų įrašui.

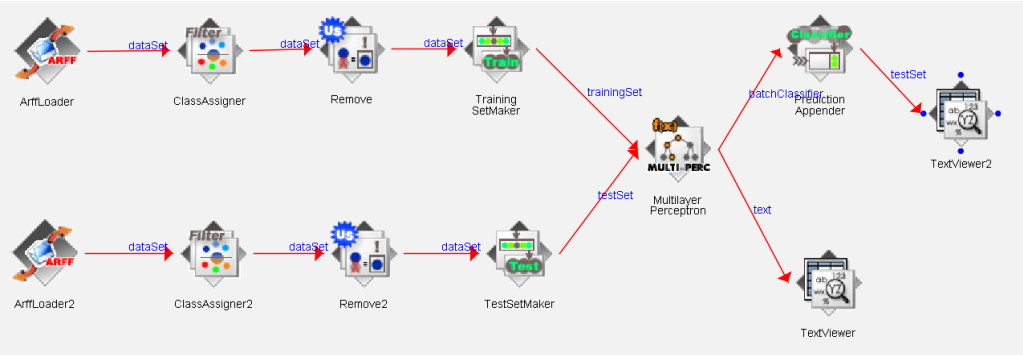
Be to, klasifikavimo tikslumo metrikos vaizduojamos 7 paveikslėlyje. Modelis teisingai priskyrė klases 30 iš 30 duomenų įrašų. Kappa statistikos reikšmė parodo, kad modelio klasifikavimas yra statistiškai reikšmingai geresnis nei atsitiktinis klasių priskyrimas.

```
=== Evaluation result ===

Scheme: MultilayerPerceptron
Options: -L 0.6 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H "4, 8" -batch-size 10
Relation: iris-weka.filters.supervised.instance.StratifiedRemoveFolds-S1-N5-F4-weka.filters.unsupervised

=== Summary ===

Correctly Classified Instances          30               100      %
Incorrectly Classified Instances         0                 0      %
Total Number of Instances               30
```

7 pav.: Klasifikavimo tikslumo metrikos.

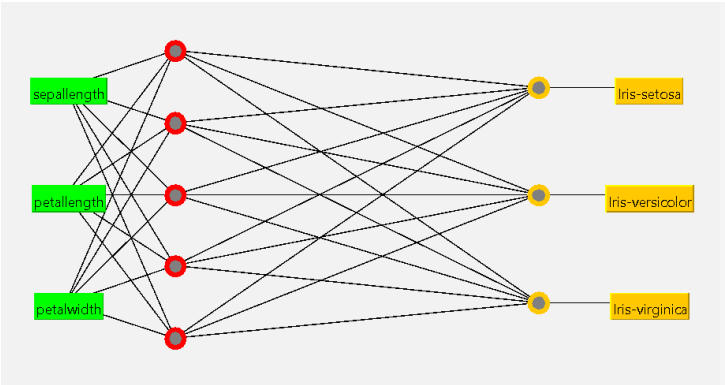
3.3 Trečioji seka

Trečiojoje užduočių sekoje naudojamas vieno paslėpto sluoksnio su 5 perceptronais neuroninis tinklas. Mokymosi greitis nustatytas 0,3 ir momentum 0,2. Šioje sekoje naudojamos mokymosi ir testavimo duomenų aibės. Užduočių sekos konstrukcija vaizduojama 8 paveiksle.



8 pav.: Trečioji užduočių seka

Gauto neuroninio tinklo su vienu paslėptu sluoksniu, kuriame yra 5 neuronai, struktūra pateikiama 9 paveiksle. Remiantis šia struktūra, gauname visų jungčių svorius, svoriai pateikiami 3 ir 4 lentelėse.



9 pav.: Neuroninio tinklo struktūra.

Node	Poslinkis	Sepallength	Petallength	Petalwidth
3	-3,156	-2,619	10,488	2,748
4	2,189	1,778	-7,549	-1,942
5	-3,871	0,026	-4,789	-4,240
6	1,257	1,007	-4,335	-1,281
7	1,931	-0,838	3,006	2,715

3 lentelė: Jungčių tarp įvesties ir paslėpto sluoksnio neuronų svoriai

Node	Poslinkis	w_0	w_1	w_2	w_3	w_4
0	-1,808	-2,828	0,877	5,386	0,771	-5,237
1	-0,260	-7,486	3,702	-9,296	1,118	3,102
2	-1,717	5,111	-5,189	-3,167	-3,718	2,201

4 lentelė: Jungčių tarp paslėpto sluoksnio neuronų ir išėjimų svoriai

4 4 Excel skaičiavimai

4.1 Formulės

Naudojant testavimo duomenis ir trečioje užduočių sekoje naudojamus svorius, atkartojame WEKA skaičiavimus Excel programoje. WEKA sistemoje įėjimo duomenys pakeičiami taip, kad jie būtų intervale $[-1, 1]$. Todėl kiekvieno požymio reikšmė normuojama naudojant formulę:

$$x_{ij} = \frac{2x_{ij} - \min(x_{1j}, x_{2j}, \ldots, x_{mj}) - \max(x_{1j}, x_{2j}, \ldots, x_{mj})}{\max(x_{1j}, x_{2j}, \ldots, x_{mj}) - \min(x_{1j}, x_{2j}, \ldots, x_{mj})}$$

Tegu f – sigmoidinė funkcija, apibrėžiama taip:

$$f(a_j) = \frac{1}{1 + e^{-a_j}}$$

Tada skaičiuojame paslėpto sluoksnio j-ojo neurono išėjimo reikšmę y_j:

$y_j = f(a_j) = f\left(\sum_{k=0}^{2} w_{jk}x_k\right)$, čia w_{jk} yra jungties iš k-tojo įėjimo į j-ąjį neuroną svoris.

Apskaičiuojame neuroninio tinklo išėjimus pagal formulę:

$$y_j = f(a_j) = f\left(\sum_{k=0}^{5} w_{jk}y_k\right)$$

Čia w_{jk} – svoriai jungčių, kurios jungia k-ąjį neuroną paslėptame sluoksnyje su j-uoju neuronu išėjimo sluoksnyje, o y_k yra paslėpto sluoksnio k-ojo neurono išėjimo reikšmė.

4.2 Excel ir WEKA rezultatų palyginimas

Suskaičiavus duomenų priskyrimo klasėms tikimybes per Excel, galima atlikti palyginimą tarp WEKA gautų reikšmių. Tikimybių palyginimas pateiktas 5 lentelėje. $P(A)_E$ žymi tikimybę įgyti klasę A pagal Excel rezultatus, o $P(A)_W$ - ši tikimybė pagal WEKA. Didžioji dalis įrašų klasifikuojami taip pat, tačiau 3 įrašai priskiriami skirtingoms klasėms, du įrašai iš Virginica priskirti Versicolor ir vienas iš Versicolor į Virginica. .

$P(Setosa)_E$	$P(Versicolor)_E$	$P(Virginica)_E$	$P(Setosa)_W$	$P(Versicolor)_W$	$P(Virginica)_W$
0,000054	0,010383	0,995326	0,000062	0,013893	0,986044
0,000062	0,013087	0,992607	0,000101	0,042482	0,957417
0,000073	0,019067	0,987134	0,000177	0,164001	0,835822
0,000077	0,021641	0,984738	0,000209	0,233681	0,766110
0,000056	0,010742	0,994864	0,000069	0,016731	0,983199
0,000060	0,011424	0,993891	0,000088	0,024642	0,975270
0,000056	0,010534	0,995110	0,000066	0,014949	0,984986
0,000054	0,010123	0,995613	0,000059	0,012299	0,987642
0,000064	0,014897	0,991431	0,000108	0,058831	0,941060
0,000061	0,012559	0,993188	0,000091	0,033892	0,966017
0,991794	0,012148	0,000001	0,988629	0,011370	0,000001
0,991144	0,013206	0,000001	0,987862	0,012137	0,000001
0,990276	0,013637	0,000001	0,987574	0,012424	0,000001
0,989786	0,015162	0,000001	0,986516	0,013482	0,000001
0,992504	0,011122	0,000001	0,989287	0,010712	0,000001
0,989538	0,016612	0,000001	0,985531	0,014468	0,000001
0,987742	0,018222	0,000001	0,984516	0,015482	0,000001
0,990909	0,012894	0,000001	0,988061	0,011938	0,000001
0,989807	0,014677	0,000001	0,986927	0,013072	0,000001
0,989261	0,015144	0,000001	0,986509	0,013490	0,000001
0,001047	0,922629	0,062192	0,004487	0,993684	0,001829
0,003837	0,992174	0,003622	0,011013	0,988536	0,000450
0,000153	0,093604	0,906927	0,000929	0,915376	0,083695
0,000261	0,311486	0,684105	0,001587	0,978890*	0,019523*
0,012799	0,997164	0,000399	0,031901	0,967960	0,000139
0,000976	0,910288	0,074111	0,004262	0,993680	0,002058
0,000203	0,176735	0,817184	0,001313	0,963299*	0,035388*
0,017489	0,996932	0,000235	0,044865	0,955039	0,000096
0,000579	0,758213	0,236072	0,002762	0,991822	0,005416
0,000103	0,039481	0,966146	0,000457	0,647073*	0,352470*

5 lentelė: Excel ir WEKA klasių priskyrimo tikimybių palyginimas.

Pastaba: Asteriksas * nurodo skirtumą tarp priskirtos klasės.

5 Išvados

Praktinės užduoties metu buvo atliktas neuroninio tinklo klasifikavimo tyrimas naudojant WEKA sistemą su „Irisų“ duomenų rinkiniu. Klasifikavimas buvo atliktas remiantis trimis požymiais: sepallength, petalwidth ir petallength, ir duomenys buvo padalinti į mokymo (80 %) ir testavimo (20 %) aibes. Naudojant įvairius neuroninio tinklo parametrus, buvo siekiama rasti geriausią klasifikavimo tikslumą. Pirmojoje užduotyje pasiektas 95 % tikslumas naudojant vieną paslėptą sluoksnį su 2–4 neuronais ir įvairiais mokymo greičiais bei momentum reikšmėmis. Antroje užduotyje, naudojant du paslėptus sluoksnius, tikslumas šiek tiek padidėjo ir pasiekė 95,83 % su 4 ir 8 neuronais pirmame ir antrame sluoksniuose. Trečioje užduotyje su paprastesniu tinklu, kur buvo tik vienas paslėptas sluoksnis su 5 neuronais, pasiektas 100 % tikslumas testavimo duomenims. Skaičiavimai atlikti tiek WEKA sistemoje, tiek MS Excel, kur buvo lyginamos tikimybės, dauguma klasių pagal tikimybes sutapo, tačiau buvo ir keli neatitikimai. Pagal gautus rezultatus galima teigti, kad pasirinkus tinkamus parametrus neuroninis tinklas pasiekia aukštą klasifikavimo tikslumą. Modelio klaidos dažniausiai atsirado dėl mažo duomenų kiekio testavimo aibėje arba dėl nepakankamo tinklo pritaikymo. Palyginus WEKA ir Excel rezultatus, aišku, kad abiejose sistemose atlikti skaičiavimai yra labai panašūs, tačiau skirtumų atsiranda.

Literatūra

[1] Iris dataset, https://storm.cis.fordham.edu/ gweiss/data-mining/datasets.html