VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ KATEDRA

Tiesioginio sklidimo DNT naudojant sistemą WEKA

3-oji skaitmeninio intelekto ir sprendimų priėmimų dalyko užduotis

Atliko: 4 kurso 5 grupės studentė
Gabrielė Žielytė (parašas)

Darbo vadovas: Prof., Dr. Olga Kurasova (parašas)

Vilnius – 2020

TURINYS

Duomenys

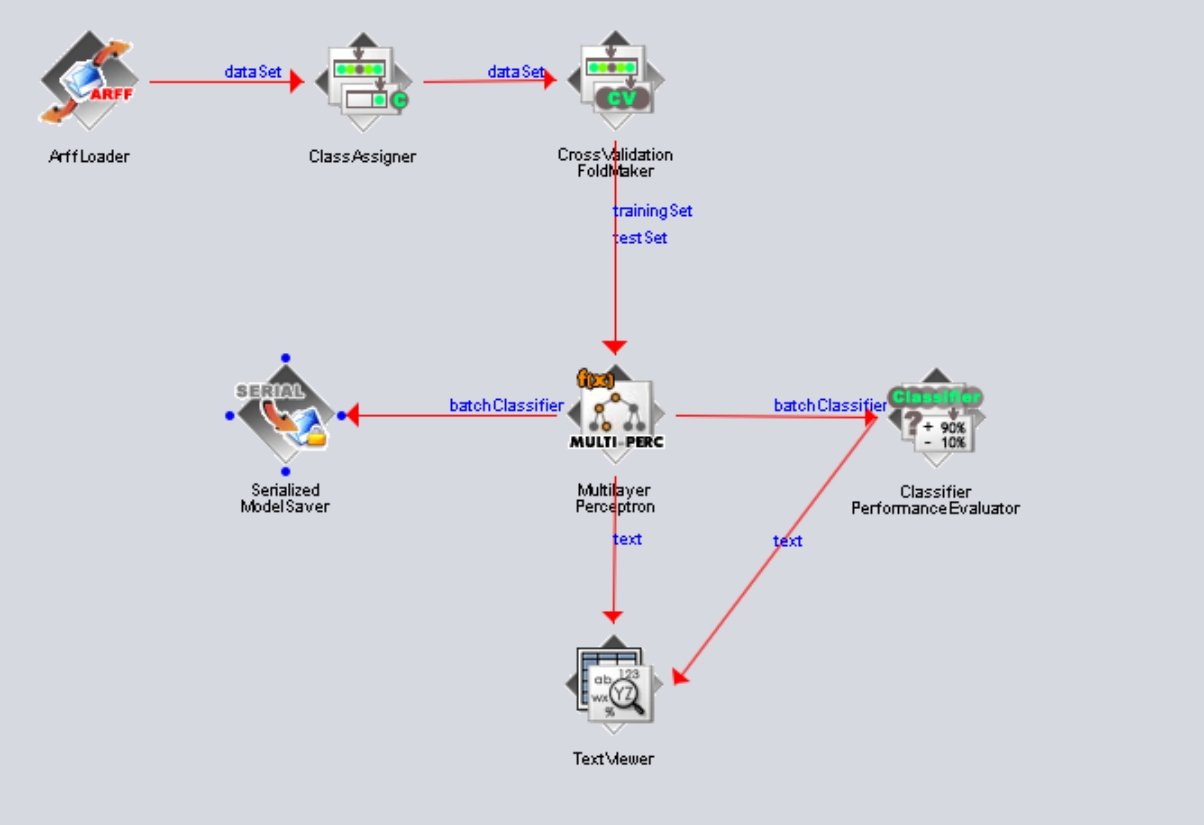
Naudojami ”irisų” duomenys, skirti klasifikuoti augalus irisus pagal 4 atributus. Yra 3 klasės - 3 irisų rūšys.

Iš failo iris.arff duomenys išskaidomi į du atskirus failus (iris_train_test.arff ir iris_new.arff). Mokymo rinkinyje kiekvienai klasei skirta 40 duomenų eilučių, testavimo - 10.

Visi duomenys yra su žinomomis klasėmis.

1. Pirmoji seka

Sistemoje WEKA sukonstruojama mokslinio darbo seka, ji įvykdoma nurodžius duomenų failą iris_train_test.arff.



1 pav. Pirmoji mokslinio darbo seka

- ArffLoader - įkrauna mokymo/testavimo duomenis iš failo.
- ArffSaver - išsaugo rezultatus .arff byloje.
- ClassAssigner - klasės atributo parinkimas, duomenys paruošiami klasifikavimui.
- Classifier PerformanceEvaluator - ištestuoja prieš tai sumodeliuoto klasifikatoriaus nuspėjimo rezultatus ”batchClassifier”, prieš tikrąsias klasių reikšmes CrossValidation FoldMaker etape sukurtų testavimo rinkinių.
- CrossValidation FoldMaker - parenkam blokų skaičių pagal kuriuos duomenų aibė bus padalinta į mokymo ir testavimo. Mūsų pasirinktas blokų skaičius - 10. Tai reiškia, kad visa gauta duomenų aibė (120 eilučių) bus padalinta į 10 rinkinių, kiekvienas jų - 12 eilučių dydžio. WEKA sistema sumodeliuos 10 skirtingų klasifikatorių - kiekvienas jų kaip testavimo duomenis naudos N-tąjį rinkinį (dydžio 12), o kaip mokymo - likusius N-1 rinkinių (likę 108 duomenys). Galutinis klasifikavimo modelis bus suvidurkintas iš šių 10 modelių.
- Multilayer Perceptron - neuronų tinklas, kuriuo klasifikuojami duomenys. Šiuo atveju, tai yra daugiasluoksnis perceptronas.

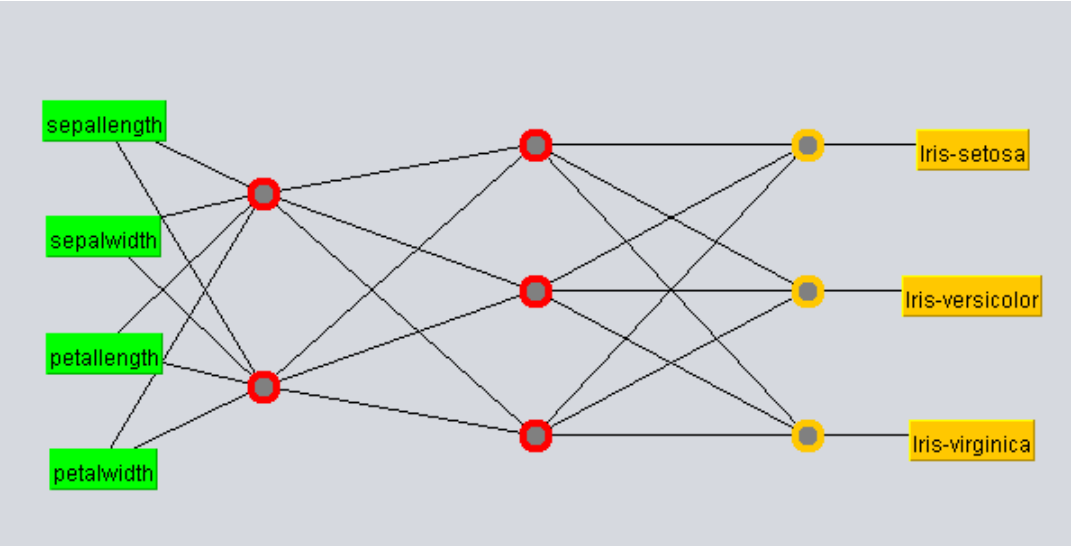
- A - Prediction Appender - klasifikatorius, kuris gauna testavimo duomenų rezultatus, kurie nebuvo naudojami mokymui.
- SerializedModelSaver - neuronų tinklui išsaugoti.
- TextViewer - rezultatams peržiūrėti.
- TextSaver - išsaugo rezultatus į teksto bylą.

1.1. Pirmasis bandymas

Kad tinklas geriausiai išmoktų klasifikuoti duomenis, naudotos tokios reikšmės:

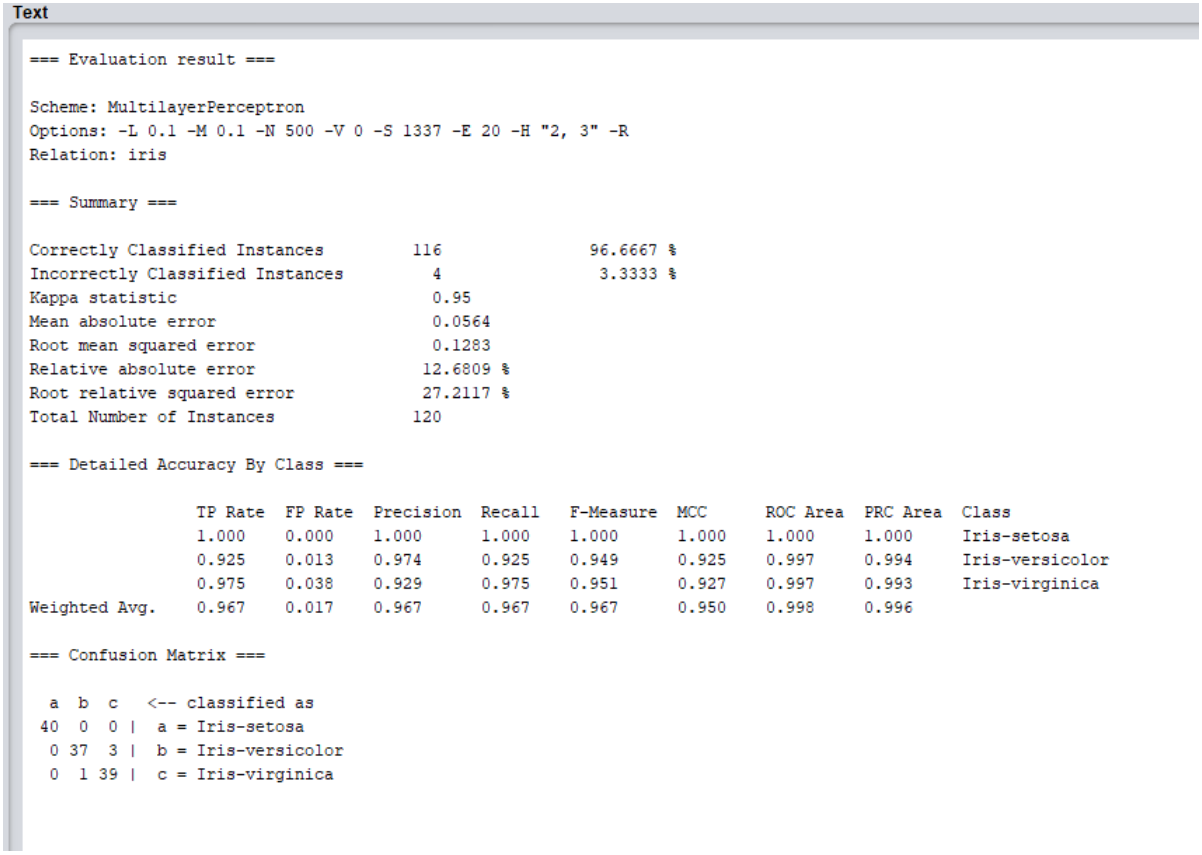
- Mokymo greitis = 0.1
- Atsparumas krypties pokyčiui (momentum) = 0.1
- Epochų skaičius = 500
- Neuronų skaičius: (2, 3) - iš viso 5 neuronai dvejuose sluoksniuose

Taip pat naudotas toks neuronų tinklas (2 pav.):



2 pav. Sekoje naudotas neuronų tinklas (1 bandymas)

TextViewer informacija:

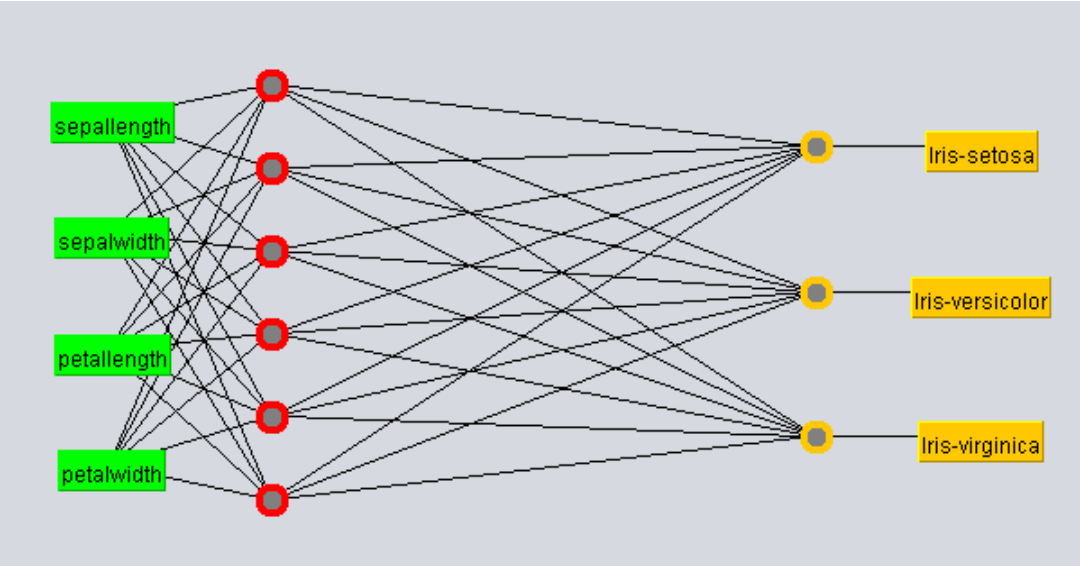


3 pav. TextViewer informacija 1

Gaunami pakankamai geri rezultatai (96.6% tikslumo) su pakankamai mažu paslėptų neuronų skaičiumi (5), tačiau įmanoma gauti ir geresnį rezultatą, kuris gaunamas antrojo bandymo metu.

1.2. Antrasis bandymas

- Mokymo greitis = 0.1
- Atsparumas krypties pokyčiui (momentum) = 0.1
- Epochų skaičius = 400
- Neuronų skaičius: 6 neuronai viename sluoksnyje



4 pav. Sekose naudotas neuronų tinklas (2 bandymas)

TextViewer informacija:

```
=== Evaluation result ===

Scheme: MultilayerPerceptron
Options: -L 0.1 -M 0.1 -N 400 -V 0 -S 1337 -E 20 -H 6 -G -R
Relation: iris

=== Summary ===

Correctly Classified Instances         117               97.5    %
Incorrectly Classified Instances         3                2.5    %
Kappa statistic                          0.9625
Mean absolute error                      0.0484
Root mean squared error                  0.1275
Relative absolute error                 10.8949 %
Root relative squared error             27.0549 %
Total Number of Instances              120

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 0.950    0.013    0.974      0.950    0.962      0.944    0.996     0.992     Iris-versicolor
                 0.975    0.025    0.951      0.975    0.963      0.944    0.996     0.991     Iris-virginica
Weighted Avg.    0.975    0.013    0.975      0.975    0.975      0.963    0.997     0.994

=== Confusion Matrix ===

  a  b  c   <-- classified as
 40  0  0 |  a = Iris-setosa
  0 38  2 |  b = Iris-versicolor
  0  1 39 |  c = Iris-virginica
```

5 pav. TextViewer informacija 2

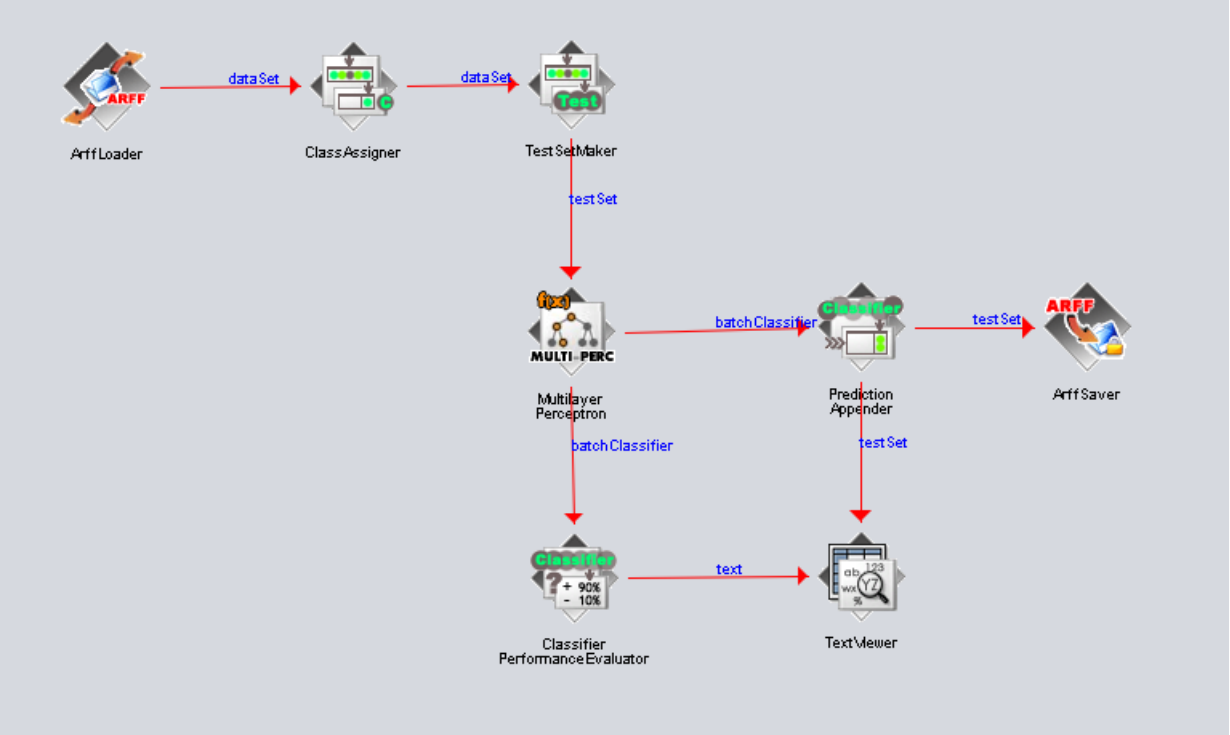
1.3. Išvados

Panaudojus vienu daugiau neuronų ir naudojant vieną neuronų sluoksnį rezultatai yra geresni, net, kai epochų skaičius yra mažesnis.

100% tikslumo pasiekti nepavyko.

2. Antroji seka

Naujų duomenų klasifikavimo seka:



6 pav. Antroji mokslinio darbo seka

Naudojame prieš tai išsaugotą perceptrono modelį, ir testuojame su naujais duomenimis iš iris_new.arff. Nauji sekos komponentai:

- A - Prediction Appender - klasifikatorius, kuris gauna testavimo duomenų rezultatus, kurie nebuvo naudojami mokymui.
- SerializedModelSaver - neuronų tinklui išsaugoti.
- TextViewer - rezultatams peržiūrėti.
- TextSaver - išsaugo rezultatus į teksto bylą.
- TestSetMaker - iš duomenų padaro testavimo aibę.
- TrainingSetMaker - iš duomenų padaroma mokymo aibė.

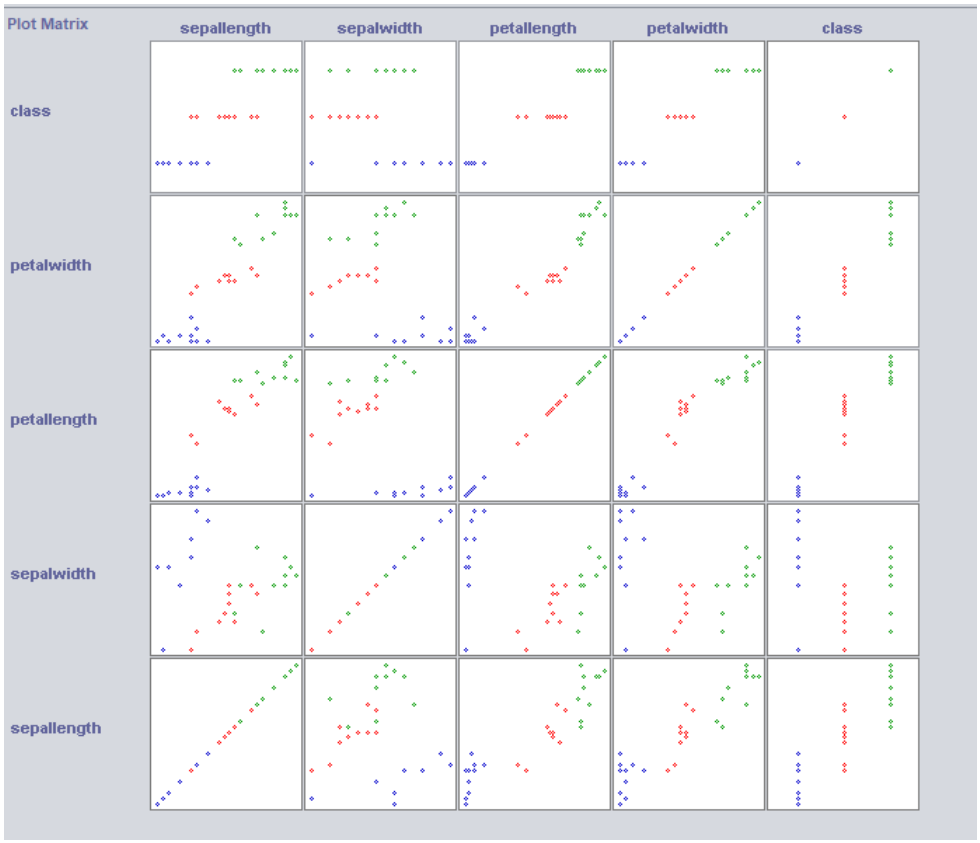
Gautas tikslumas: 100%. Buvo naudoti pirmosios sekos antrojo bandymo parametrai. TextViewer informacija:

Text

```
=== Evaluation result ===

Scheme: MultilayerPerceptron
Options: -L 0.3 -M 0.1 -N 400 -V 0 -S 1337 -E 20 -H "2, 4" -G -R
Relation: iris

=== Summary ===

Correctly Classified Instances          30               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0332
Root mean squared error                  0.0921
Total Number of Instances               30

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-setosa
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-versicolor
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-virginica
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b  c   <-- classified as
 10  0  0 |  a = Iris-setosa
  0 10  0 |  b = Iris-versicolor
  0  0 10 |  c = Iris-virginica
```

7 pav. TextViewer informacija antrai sekai

Duomenų požymių (stulpelių) porų vaizdai Dekarto koordinačių sistemoje (Scatter Plot Matrix):



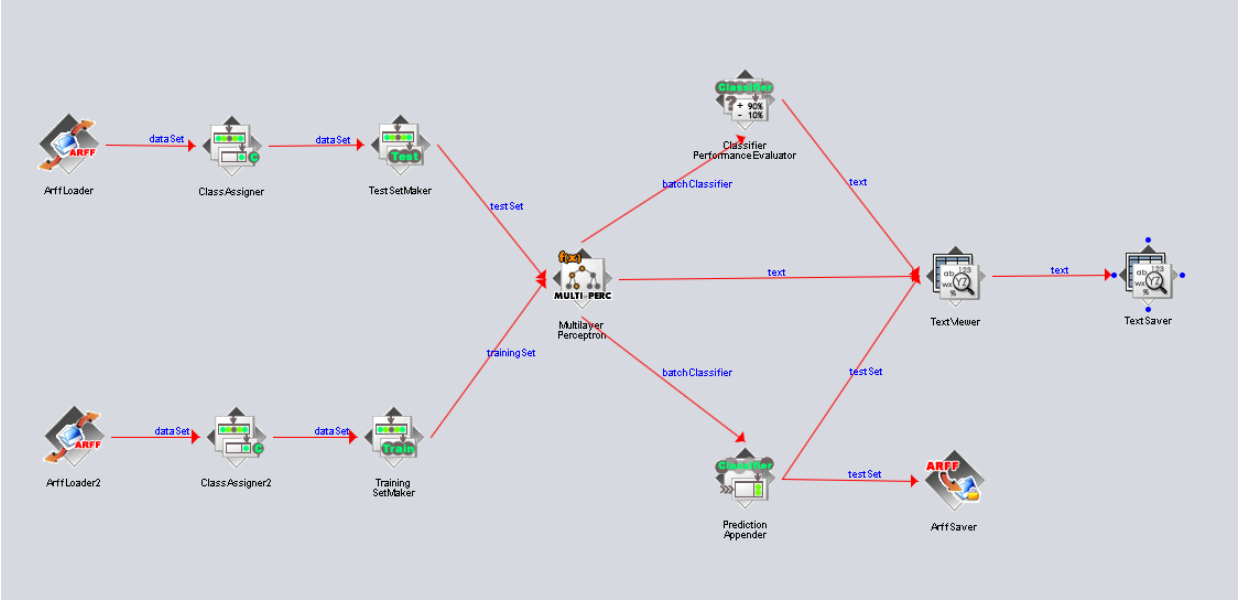
8 pav. Scatter Plot Matrix

3. Trečioji seka - klasifikavimas ir testavimas

Atskirai pakraunamos mokymo ir testavimo duomenų aibės. Rezultatai išsaugomi. Gautas 100% tikslumas.

Parametrai:

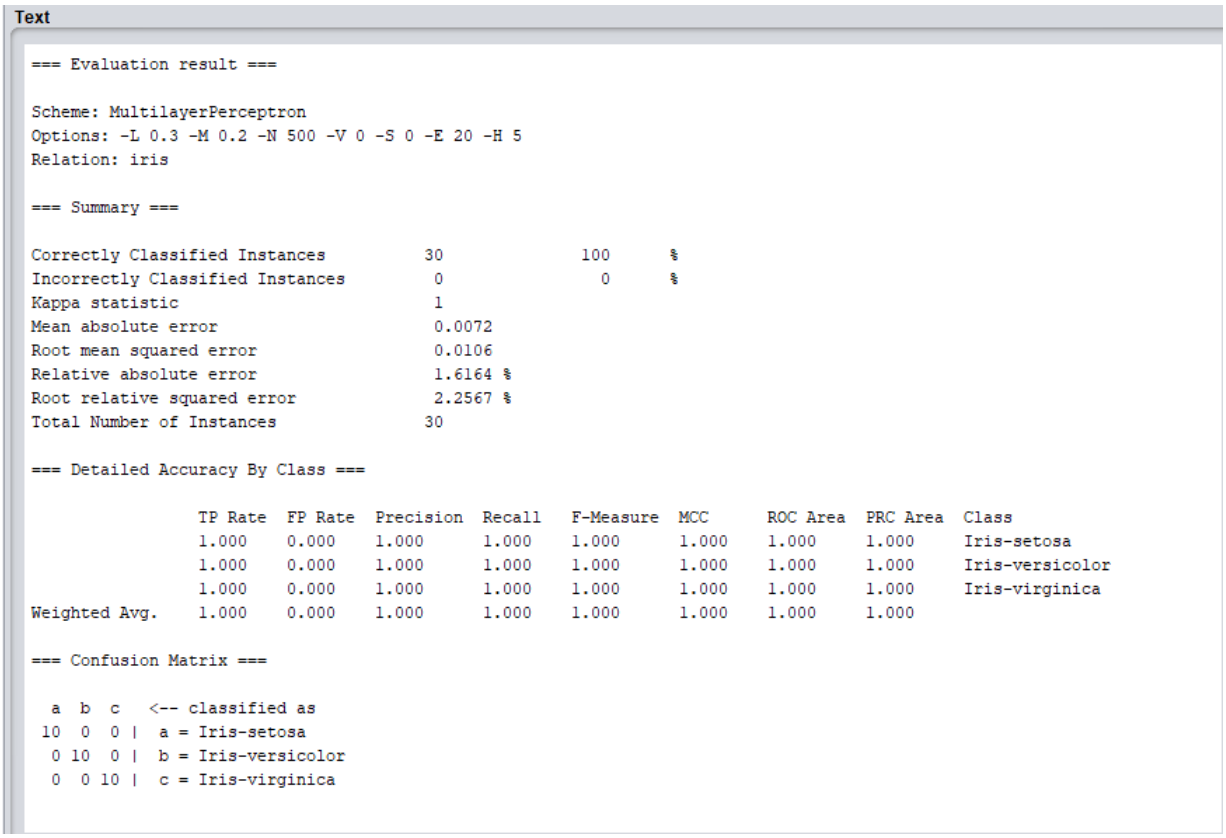
- Mokymo greitis = 0.1
- Atsparumas krypties pokyčiui (momentum) = 0.2
- Epochų skaičius = 500
- Neuronų skaičius: 6 neuronai viename sluoksnyje



9 pav. Trečioji seka

Neuroninio tinklo svoriai ir testavimo duomenų priskyrimas klasėms išsaugomas faile savedText.txt.

TextViewer informacija:



10 pav. TextViewer informacija trečiajai sekai

4. Neuronų išėjimo reikšmių perskaičiavimas MS Excel programoje

Sudaromas neuronų tinklas skaičiuoklėje. Rezultatų palyginimas:

Wekos reikšmės			Skaičiuoklės reikšmės			Skirtumas		
0.981708	0.018291	0.000001	0.98759318	0.01582652	7.6264E-07	0.00588518	0.00246448	2.3736E-07
0.934023	0.065975	0.000002	0.94792178	0.07008124	2.205E-06	0.01389878	0.00410624	2.0498E-07
0.981203	0.018796	0.000001	0.98714836	0.01626718	7.9108E-07	0.00594536	0.00252882	2.0892E-07
0.974307	0.025692	0.000001	0.98764177	0.01784668	8.5959E-07	0.01333477	0.00784532	1.4041E-07
0.980427	0.019572	0.000001	0.98746548	0.0159752	7.8101E-07	0.00703848	0.0035968	2.1899E-07
0.974545	0.025454	0.000001	0.98447105	0.01966094	9.1162E-07	0.00992605	0.00579306	8.8382E-08
0.983075	0.016924	0.000001	0.98800574	0.015347	7.4112E-07	0.00493074	0.001577	2.5888E-07
0.980398	0.019601	0.000001	0.98687871	0.01661446	8.0281E-07	0.00648071	0.00298654	1.9719E-07
0.982688	0.017311	0.000001	0.98790376	0.01548275	7.4372E-07	0.00521576	0.00182825	2.5628E-07
0.98064	0.019359	0.000001	0.98709162	0.01643225	7.8625E-07	0.00645162	0.00292675	2.1375E-07
0.011393	0.982847	0.00576	0.00580806	0.94447995	0.0332192	0.00558494	0.03836705	0.0274592
0.010058	0.981249	0.008693	0.00677234	0.95653705	0.02360093	0.00328566	0.02471195	0.01490793
0.017678	0.980812	0.001509	0.01242594	0.99068979	0.00317015	0.00525206	0.00987779	0.00166115
0.033963	0.96559	0.000447	0.02112516	0.99430376	0.00087442	0.01283784	0.02871376	0.00042742
0.012957	0.982954	0.004089	0.00804996	0.97254526	0.01337021	0.00490704	0.01040874	0.00928121
0.024578	0.974331	0.001091	0.01954159	0.9900925	0.00170492	0.00503641	0.0157615	0.00061392
0.016913	0.980781	0.002306	0.01239783	0.98543042	0.00461714	0.00451517	0.00464942	0.00231114
0.015918	0.981963	0.002119	0.01234087	0.9893375	0.00360321	0.00357713	0.0073745	0.00148421
0.050611	0.949122	0.000267	0.03519192	0.99312607	0.00040335	0.01541908	0.04400407	0.00013635
0.016496	0.981281	0.002224	0.01184639	0.98649228	0.00462972	0.00464961	0.00521128	0.00240572
0.000222	0.003148	0.99663	0.00019602	0.00235322	0.99913721	2.5979E-05	0.00079478	0.00250721
0.000356	0.0107	0.988944	0.00027806	0.00561056	0.99711248	7.794E-05	0.00508944	0.00816848
0.000422	0.016364	0.983215	0.00025589	0.00451011	0.99779941	0.00016611	0.01185389	0.01458441
0.000223	0.003158	0.996619	0.00019604	0.00233822	0.99913969	2.6963E-05	0.00081978	0.00252069
0.000212	0.002789	0.996999	0.00019313	0.00225773	0.98593418	1.8875E-05	0.00053127	0.01106482
0.000291	0.00632	0.993389	0.00023234	0.00355045	0.99841652	5.8659E-05	0.00276955	0.00502752
0.000473	0.023698	0.975828	0.00027209	0.00547241	0.99718203	0.00020091	0.01822559	0.02135403
0.000505	0.026404	0.973091	0.00032347	0.00823408	0.99527735	0.00018153	0.01816992	0.02218635
0.000286	0.00553	0.994185	0.00023645	0.0034309	0.99844447	4.9552E-05	0.0020991	0.00425947
0.000835	0.087684	0.911481	0.00043306	0.01665156	0.98878376	0.00040194	0.07103244	0.07730276
					average	0.00484603	0.01187064	0.00765553

11 pav. Rezultatų palyginimas

Gauti svoriai ir visi tarpiniai skaičiavimai yra atskirame skaičiuoklės faile finalNet.xlsx .

Galima matyti, kad skirtumai tarp skaičiuoklės ir Wekos reikšmių nebuvo pernelyg dideli, todėl parametrai buvo pakankamai gerai pasirinkti, ir Wekos programa gana gerai klasifikavo duomenis, bet tikriausiai sutrukdė Wekos verčių apvalinimas, nes rezultatai nėra visiškai tikslūs.