

#### IV užduotis (Tiesioginio sklidimo DNT naudojant sistemą WEKA)

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

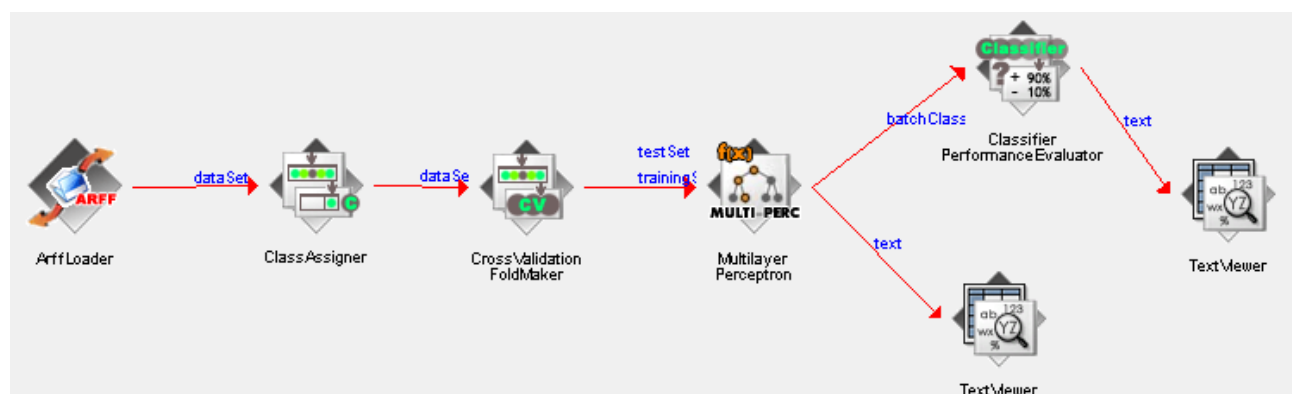
##### **1) Duomenų paruošimas**

Šiame darbe bus naudojami irisų arba kiti norimi duomenys. Irisų duomenų *arff* failas įrašomas į kompiuterį įdiegus sistemą WEKA. Galima naudoti ir kitus įrašytus duomenis arba susirasti patiems, pavyzdžiui saugykloje <https://archive.ics.uci.edu/ml/datasets.html>

Iš šio failo reikia padaryti du failus: *iris\_train\_test.arff* ir *iris\_new.arff*. Pirmajame palikti po 40 kiekvienos klasės duomenų, o antrajame – likę 10 (kiekvienai klasei). Be to, galima ištrinti failo pradžioje nurodytus komentarus.

##### **2) Mokslinio darbo sekos sukonstravimas**

Sistemoje WEKA sukonstruokite paveiksluke pateiktą mokslinio darbo seką. Ją įvykdysite nurodžius duomenų failą *iris\_train\_test.arff*.



##### **3) Neuroninio tinklo parametrų parinkimas**

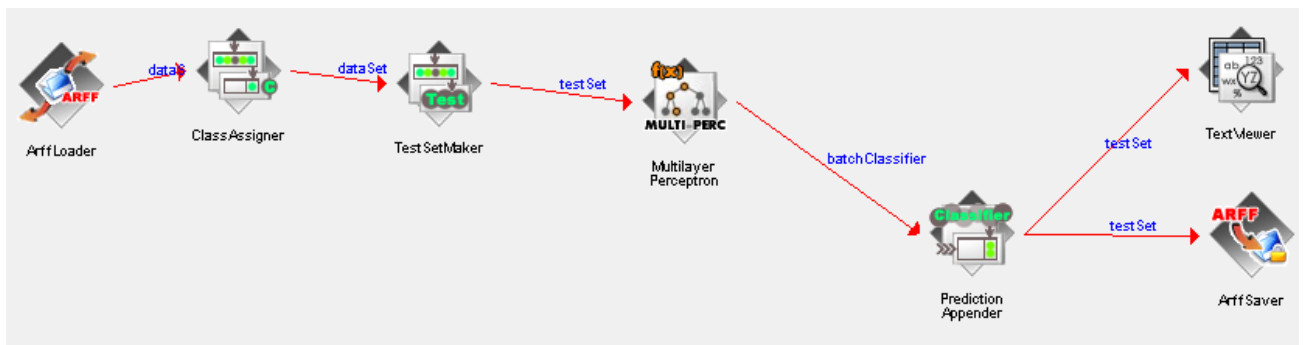
Parinkite tokius paslėptų neuronų skaičius, mokymo greičio parametro bei *momentum* reikšmes, kad tinklas geriausiai išmokyti klasifikuoti duomenis. Klasifikavimo tikslumą vertinkite pagal teisingai klasifikuotų duomenų kiekį.

Išsaugokite tinklo modelį, pagal kurį gaunami tiksliausi klasifikavimo rezultatai.

##### **4) Naujų duomenų klasifikavimas**

Sukurkite ir įvykdysite dar vieną mokslinio darbo seką, kad nauji duomenys su nežinomomis klasėmis būtų priskirti klasėms (naudokite failą *iris\_new.arff*) pagal sukurtą ir išsaugotą tinklo modelį.

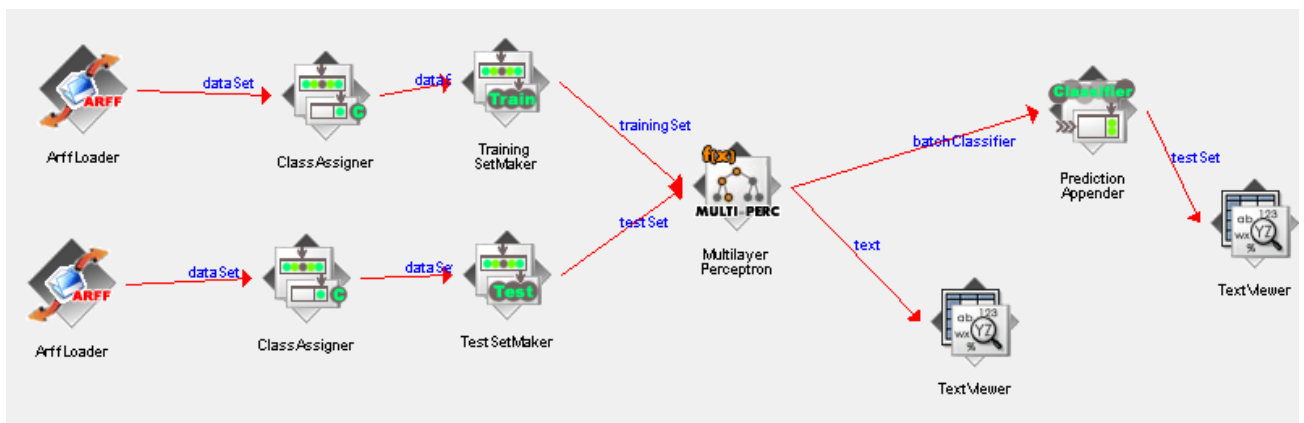
Išsaugokite gautus rezultatus.



## 5) Klasifikavimas ir testavimas

Sukurkite ir įvykdykite paveikslėlyje pateiktą mokslinio darbo seką (mokymo duomenys *iris\_train\_test.arff*, testavimo *iris\_new.arff*).

Nustatykite tik vieno sluoksnio paslėptų neuronų skaičių (pasirinkite iš galimų variantų 5, 6 ar 7).



Išsaugokite abiejų komponentų TextViewer pateiktus rezultatus.

## 6) Neuronų išėjimo reikšmių perskaiciavimas MS Excel programoje

**Tikslas:** sukonstruoti neuroninį tinklą MS Excel aplinkoje žinant neuronų svorius, gautus sistema WEKA.

### Veiksmai atliekami MS Excel programoje:

6.1 Nauji duomenys, kurie nebuvo naudojami neuroniniam tinklui mokyti, su tinklo priskirtų klasių tikimybėmis iš TextViewer nukopijuojami į MS Excel lentelę.

Kadangi WEKA sistemoje skaičiaus sveikąją dalį nuo trupmeninės skiria taškas, o MS Excel – kablelis (lietuvių k.), tai prieš kopijuojant duomenis reikia tuo pasirūpinti (kablelius pakeisti į tarpus, o taškus – į kablelius)

6.2 WEKA sistemoje, jeigu nenustatyta kitaip, įėjimo duomenys pakeičiami taip, kad jei būtų intervale  $[-1 \ 1]$ . Todėl reikia į MS Excel lentelę įkeltus duomenis „suvesti“ į šį intervalą.

Tegu turime duomenis  $X_1, X_2, \dots, X_m$ , ( $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ ,  $i = 1, \dots, m$ ) norint pakeisti jų požymių reikšmių mastelį, pavyzdžiui į  $[-1, 1]$ , t. y., kad mažiausia reikšmė būtų  $-1$ , didžiausia  $1$ , atliekama transformacija vadinama normavimu.

$$x_{ij} \leftarrow \frac{2x_{ij} - \min_{(x_{1j}, x_{2j}, \dots, x_{mj})} - \max_{(x_{1j}, x_{2j}, \dots, x_{mj})}}{\max_{(x_{1j}, x_{2j}, \dots, x_{mj})} - \min_{(x_{1j}, x_{2j}, \dots, x_{mj})}}$$

- 6.3 Perkelkite neuronų svorių lenteles, gautas 5 žingsnyje į MS Excel lentelę (turi būti dvi lentelės).
- 6.4 Susumuokite duomenų įėjimo vektorių ir paslėptų neuronų svorių vektorių sandaugas ( $a_1 = \sum_{k=0}^n w_{1k} x_k$  ir  $a_2 = \sum_{k=0}^n w_{2k} x_k$ ) esant dviem paslėptiems neuronams, esant daugiau paslėptų neuronų atitinkamas kiekis turi būti ir šių sumų  $a_j$ ) (šias sandaugas reikia apskaičiuoti visiems duomenų įėjimo vektoriams).
- 6.5 Apskaičiuokite sigmoidinės funkcijos reikšmes ( $f(a_1) = \frac{1}{1+e^{-a_1}}$  ir  $f(a_2) = \frac{1}{1+e^{-a_2}}$ ) nuo gautų sumų. Esant daugiau paslėptų neuronų atitinkamas kiekis turi būti ir šių funkcijų reikšmių  $f(a_j)$  (šias funkcijų reikšmes reikia apskaičiuoti visiems duomenų įėjimo vektoriams).
- 6.6 Susumuokite 6.5 punkte gautų funkcijų reikšmių vektorių ir paslėptų neuronų svorių vektorių sandaugas (šias sandaugas reikia apskaičiuoti visiems duomenų įėjimo vektoriams).
- 6.7 Apskaičiuokite sigmoidinės funkcijos reikšmes nuo gautų sumų (šias funkcijų reikšmes reikia apskaičiuoti visiems duomenų įėjimo vektoriams).
- 6.8 Trijų klasių atveju rezultate turi gautis trys stulpeliai, parodantys tikimybes, pagal kurias duomenys priskiriami klasei su didžiausia tikimybe. Šios tikimybės yra suskaičiuotos ir neuroninio tinklo. Sulyginkite gautus rezultatus (jie tam tikru tikslumu turi sutapti).

### Užduoties ataskaitoje:

- Trumpai apibūdinkite analizuojamus duomenis, kiek jų yra naudota tinklui mokytis ir testuoti, kiek duomenų su nežinomomis klasėmis;
- Pateikite sudarytas mokslinio darbo sekas. Pateikite komentarus, ką reiškia kiekviena naudojama komponentė.
- Nurodykite, kiek turi būti paslėptų neuronų, kokios mokymo greičio parametro bei *momentum* reikšmės, kad tinklas geriausiai išmoktų klasifikuoti duomenis. Pateikite klasifikavimo rezultatus (informacija iš TextViewer).
- Pateikite neuroninio tinklo vaizdą.
- Padarykite išvadą apie naujų duomenų, kurių klasės nežinomos, klasifikavimo rezultatus.

- Pateikite duomenų porų vaizdus Dekarto koordinatų sistemoje (Scatter Plot Matrix).
- Pateikite 5 žingsnyje gautų neuronų svorių reikšmes, surašytas į lentelę.
- MS Excel programa gautus rezultatus; aprašyti kaip buvo konstruojamas neuroninis tinklas.

**P. S.** Kiekvieną atliktą žingsnį trumpai pakomentuoti, įvardinti visus pateiktus žymėjimus.