III 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/index.php.

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.

SVARBU: Būtina klasifikuoti irisų duomenis ne pagal keturis požymius (*features, attributes*), bet tik pagal tris. Pagal studento numerį reikia pasirinkti vieną iš variantų (studento numerio paskutinio skaitmens dalybos iš trijų liekana).

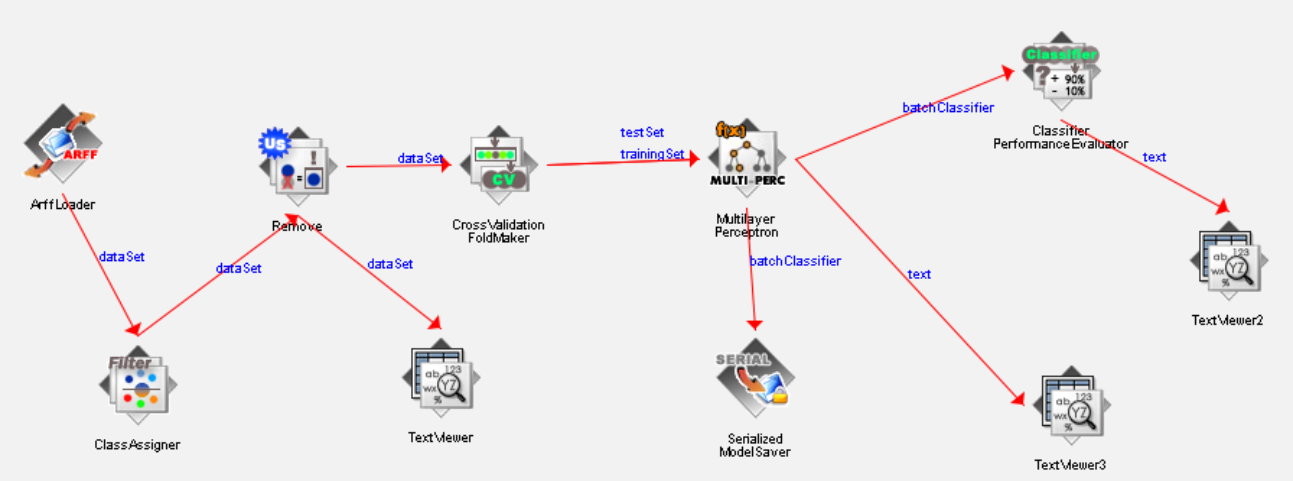
Variantai:

0. sepallength, sepalwidth, petallength
1. sepalwidth, petallength, petalwidth
2. sepallength, petallength, petalwidth

2) Pirmos užduočių sekos sukonstravimas daugiasluoksniam perceptronui apmokyti

Sistemoje WEKA sukonstruokite 1 paveiksliuke pateiktą užduočių seką:

- Ją įvykdykite nurodžius duomenų failą *iris_train_test.arff*.
- Komponentėje *Remove* nurodykite požymio (atributo) indeksą, kurį norite išmesti.
- Komponentėje *SerializedModelSaver* nurodykite kompiuterio vietą (aplanką), kurioje bus išsaugotas išmokytas modelis.
- Komponentėje *CrossValidation FoldMaker* kryžminės patikros blokų (*Number of folds*) skaičių pakeiskite į 5.
- Komponentėje *MultiLayer Perceptron* paketo dydį *batchSize* pakeiskite į 10.



1 pav. Pirma užduočių seka

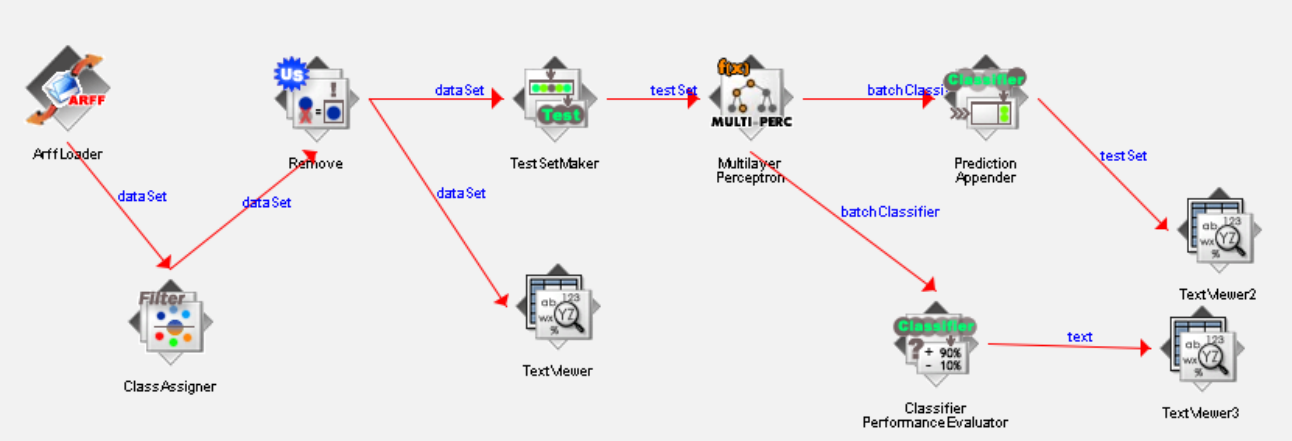
3) Neuroninio tinklo parametrų parinkimas

Komponentėje *MultiLayer Perceptron* parinkite tokius paslėptų neuronų skaičius (*hiddenLayers*), mokymo greičio parametro (*learningRate*) bei *momentum* reikšmes, kad tinklas geriausiai išmoktų klasifikuoti duomenis. Klasifikavimo tikslumą vertinkite pagal teisingai klasifikuotų duomenų kiekį (žr. komponentę *Classifier Performance* → *TextViewer*).

P. S. WEKA sistemoje nustatyta *hiddenLayers* reikšmė „a“. Norint sukurti neuroninį tinklą iš kelių paslėptų sluoksnių, reikia nurodyti kiekvieno sluoksnio neuronų skaičių (jei turi būti atskirti kableliais, pvz., 3,5).

4) Antros užduočių sekos konstravimas naujiems duomenims klasifikuoti

Sukurkite ir įvykdykite dar vieną užduočių seką (žr. 2 pav.), kad nauji duomenys su nežinomomis klasėmis būtų priskirti klasėms (naudokite failą *iris_new.arff*) pagal sukurtą ir išsaugotą tinklo modelį.

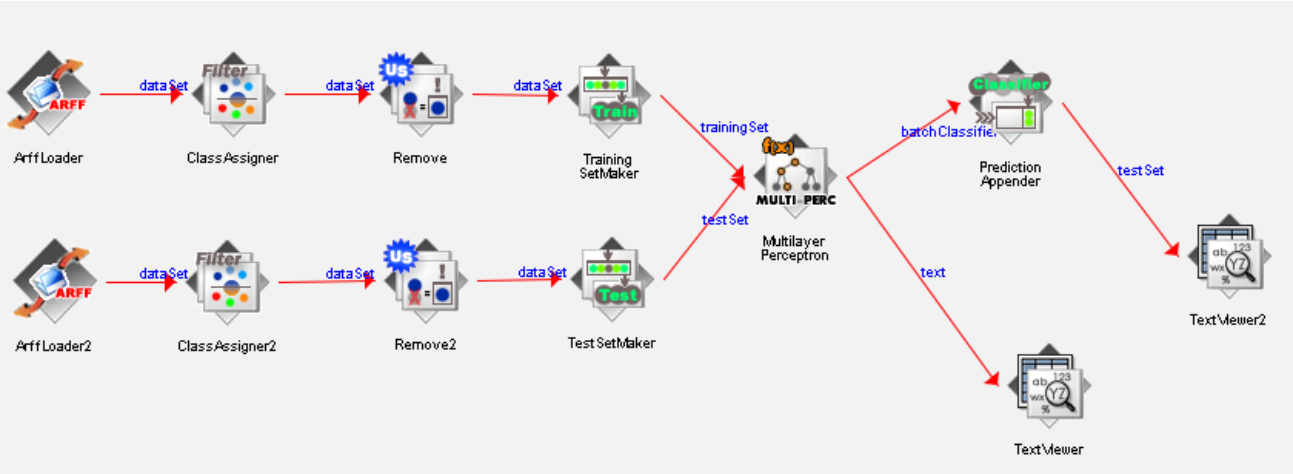


2 pav. Antra užduočių seka

5) Trečios užduočių sekos konstravimas duomenims klasifikuoti ir testuoti

Sukurkite ir įvykdykite 3 paveiksliuke pateiktą užduočių seką (mokymo duomenys *iris_train_test.arff* , testavimo *iris_new.arff*).

Nustatykite tik vieno paslėpto sluoksnio neuronų skaičių (pasirinkite iš galimų variantų 5, 6 ar 7). Komponentei *PredictionAppender* reikia nurodyti *Append Probability True*, kad galima būtų peržiūrėti ne tik kokioms klasėms duomenys yra priskirti, bet ir klasių tikimybes.



3 pav. Trečia užduočių seka

6) Neuronų išėjimo reikšmių perskaičiavimas 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.

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

P. S. Patikrinkite, koks normavimo būdas yra naudojamoje *WEKA* versijoje ir priderinkite tinkamą normavimą *MS Excel*.

6.3 Perkelkite neuronų svorių lenteles, gautas 5 žingsnyje į *MS Excel* lentelę (turi būti dvi lentelės, kadangi naudojamas vienas paslėptas sluoksnis: vienoje lentelėje svoriai jungčių tarp įvesties ir paslėpto sluoksnio neuronų, kitoje lentelėje svoriai jungčių tarp paslėpto sluoksnio neuronų ir išėjimų).

6.4 Susumuokite duomenų įėjimo vektorių ir paslėptų neuronų svorių vektorių sandaugas (pvz., jei turėtumėme tik du paslėptus neuronus, skaičiuotume $a_1 = \sum_{k=0}^{n} w_{1k}x_k$ ir $a_2 = \sum_{k=0}^{n} w_{2k}x_k$; esant daugiau paslėptų neuronų atitinkamas kiekis turi būti ir šių sumų a_j,

čia n yra duomenų požymių kiekis) (šias sandaugas reikia apskaičiuoti visiems duomenų įėjimo vektoriams).

6.5 Apskaičiuokite paslėpto sluoksnio išėjimus, t. y., sigmoidinės funkcijos reikšmes ($f(a_1) = \frac{1}{1+e^{-a_1}}$ ir $f(a_2) = \frac{1}{1+e^{-a_2}}$) nuo sumų, gautų 6.4 punkte. 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 (t. y. paslėpto sluoksnio išėjimų) vektorių ir paslėptų neuronų svorių vektorių sandaugas (šias sandaugas reikia apskaičiuoti visiems duomenų įėjimo vektoriams).

6.7 Apskaičiuokite neuroninio tinklo išėjimus, t. y., 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 sistemoje *WEKA* (komponentei *PredictionAppender* reikia nurodyti *Append Probability True*). Palyginkite gautus rezultatus, atsakant į klausimą, ar duomenys priskirti toms pačioms klasėms, kokie skirtumai yra tarp tikimybių, gautų *MS Excel* ir *WEKA*.

P. S. 6-ą punktą galima atlikti naudojant kitą programą arba suprogramavus reikiamus komponentus.

Užduoties ataskaitoje:

- Aprašykite analizuojamus duomenis, kiek jų yra naudota tinklui mokyti ir testuoti, kiek yra duomenų, kurių klasės nėra naudojamos neuroniniam tinklui mokyti, kokie duomenų požymiai (atributai) yra naudoti.
- Pateikite jūsų sudarytų užduočių sekų vaizdus (ekrano kopijas). Negalima kopijuoti į ataskaitą pateiktas užduočių sekas.
- Nurodykite, kiek turi būti paslėptų neuronų, kokios mokymo greičio parametro bei *momentum* reikšmės, kad tinklas geriausiai išmoktų klasifikuoti duomenis. Pateikite gautus klasifikavimo tikslumo įverčius (informacija iš pirmos užduočių sekos *Classifier PerformanceEvaluator* → *TextViewer*) visiems tirtiems atvejams (kurių turi būti keletas, kad galima būtų daryti apibendrintas išvadas).
- Pateikite neuroninio tinklo vaizdą. Tam reikia komponentei *MultiLayer Perceptron* nurodyti *GUI True*. Padarius neuroninio tinklo ekrano vaizdą, tolimesniems tyrimams galima naudoti *GUI False*.
- Pateikite naujų duomenų, kurių klasės nežinomos (failas *iris_new.arff*), klasifikavimo rezultatus (kad matytųsi, kokiai klasei ir su kokia tikimybe kiekvienas duomenų įrašas yra priskirtas). Informaciją imti iš antros užduočių sekos *Prediction Appender* → *TextViewer*. Taip pat pateikite klasifikavimo tikslumo metrikas (informaciją imti iš antros užduočių sekos *Classifier PerformanceEvaluator* → *TextViewer*). Padarykite išvadą apie tai, kaip gerai neuroninis tinklas klasifikavo duomenis.

- Pateikite duomenų požymių (stulpelių) porų vaizdus Dekarto koordinačių sistemoje. Tam reikia pirmoje ir antroje sekoje pridėti komponentę *ScatterPlotMatrix*, padidinkite taškų dydį *PointSize*, kad geriau jie matytųsi).
- Pateikite 5 žingsnyje gautų neuronų svorių reikšmes, surašytas į lenteles, kad būtų aišku, kurio sluoksnio kuris svorių rinkinys yra.
- *MS Excel* programa (ar kita programa) gautus rezultatus; aprašyti kaip buvo konstruojamas neuroninis tinklas. Vienoje lentelėje pateikite ir *WEKA* gautus klasifikavimo rezultatus (tikimybes), ir gautas *MS Excel* programoje. Padarykite išvadą apie rezultatų sutapimą.
- Kartu su ataskaita pateikite ir *MS Excel* failą.

P. S. Ataskaitoje turi būti aprašytas kiekvienas atliekamas veiksmas, pateikti žymėjimų aprašymai ir kita, jūsų manymu, svarbi informacija.