Akademia Górniczo-Hutnicza w Krakowie
Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

Inżynieria wiedzy i uczenie maszynowe

Konspekt zajęć laboratoryjnych
prowadzonych w Katedrze Informatyki
Studia Drugiego Stopnia
Drugi rok

Bartłomiej Śnieżyński

Laboratorium nr 4

Temat
System Weka

Cel
Celem zajęć jest wprowadzenie do systememu Weka i zapoznanie się z jego dwoma modułami: Explorer i Experimenter.

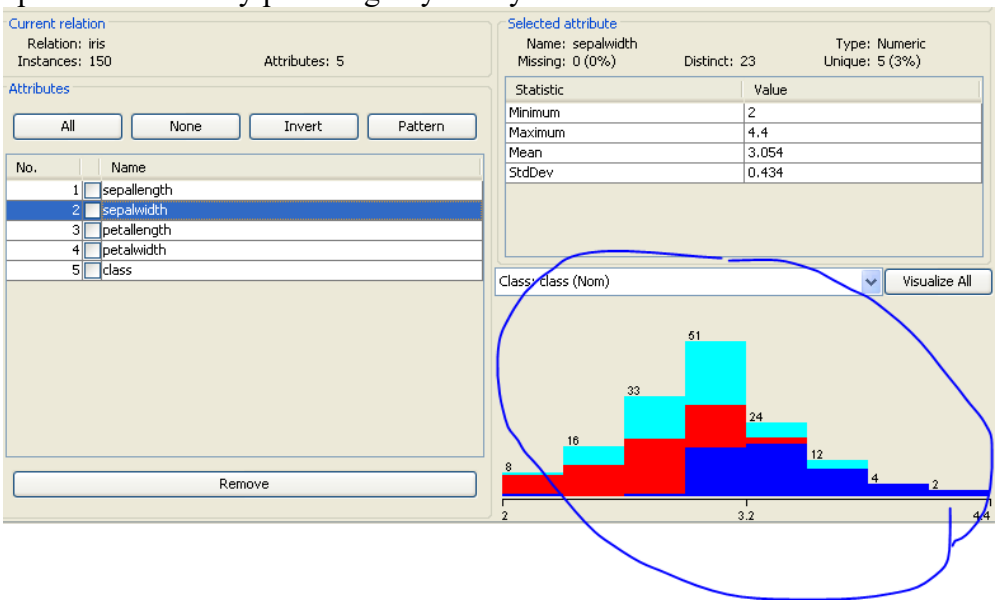
Wymagane wiadomości wstępne z wykładu
Problem klasyfikacji

Konfiguracja komputera
Podczas laboratorium wykorzystywany będzie system Weka.

Linki
http://www.cs.waikato.ac.nz/ml/weka/

Plan laboratorium

1. Uruchomić system Weka
2. Uruchomić moduł Explorer
   2.1. Otworzyć plik z danymi Iris (C:\Program Files\Weka-3-6\data)
   2.2. Sprawdzić rozkłady poszczególnych atrybutów:



   2.3. Obejrzeć wizualizację (zakładka Visualize)
      2.3.1. Przesunąć suwak PointSize oraz Jitter, a następnie wcisnąć Update. Jaki jest tego efekt?
      2.3.2. Które dwa atrybuty dobrze nadają się do klasyfikacji?
   2.4. Przejść do zakładki Classify i wygenerować kilka klasyfikatorów: z różną reprezentacją wiedzy: J48, JRip, NaiveBayes, ZeroR, BayesNet (zmienić parametr P - maxNrOfParents na liczbę 4)
   2.5. Porównać wygenerowaną wiedzę (w przypadku J48 i BayesNet można kliknąć prawym przyciskiem na elemencie Result list i wybrać VisualiseTree/Graph)
   2.6. Oglądnąć błędy klasyfikacji: PKlik na Result list/Visualise classifier erors
3. Uruchomić moduł Experimenter
   3.1. Utworzyć nową konfigurację (New)
   3.2. Dodać 5 zestawów danych (Datasets)
   3.3. Dodać 5 metod uczenia (Algorithms) z różną reprezentacją wiedzy
   3.4. Uruchomić eksperyment w zakładce Run
   3.5. Oglądnąć wyniki w zakładce Analyse (wyniki pojawią się po wciśnięciu przycisku Experiment, a następnie Perform test). V oznacza wynik statestycznie lepszy, a * statystycznie gorszy od pierwszego algorytmu:

Dataset	(1) rules.De	(2) bayes	(3) rules	(4) trees	(5) rules
iris	(100) 92.93	95.53 v	93.93	94.73	93.93
weather	(100) 64.50	67.50	67.50	66.50	36.00 *
soybean	(100) 83.97	92.94 v	91.80 v	91.78 v	39.75 *
(v/ /*)		(2/1/0)	(1/2/0)	(1/2/0)	(0/1/2)

   3.6. Zmienić output format na latex i powtórzyć test.

3.7. Powtórzyć operację dla eksperymentu typu train/test percentage split (parametr Experiment Type w zakładce Setup).

4. Stworzyć plik ARFF opisujący mieszkania do wynajęcia (metraż, liczba pokoi, wyposarzenie, koszty, piętro, winda itp.) z podziałem na klasy czy dane mieszkanie by nas interesowało czy nie.
   4.1. Zwizualizować przykłady na diagramach.
   4.2. Porównać wiedzę wygenerowaną przez różne algorytmy pod względem poprawności klasyfikacji i czytelności wiedzy.
5. Przeprowadzić operację „backward elimination" aby wybrać najlepsze atrybuty dla zbioru *glass*:
   5.1. Uruchomić uczenie przy użyciu algorytmu *IBk* dla kolejnych zestawów atrybutów, z których usunięto jeden z nich i sprawdzić dla którego poprawność klasyfikacji (z kroswalidacją) będzie największa.
   5.2. Zapisać wyniki w tabeli takiej jak poniżej.
   5.3. Procedurę wykonać rekurencyjnie wybierając do kolejnego etapu najlepszy zestaw atrybutów.

Liczba atrybutów	Atrybuty w najlepszym podzbiorze	Poprawność klasyfikacji
9		
8		
7		
6		
5		
4		
3		
2		
1		

6. **Zadanie domowe**. Wybrać dziedzinę i przygotować dla niej 20-30 przykładów treningowych. Porównać wiedzę wygenerowaną przez różne algorytmy pod względem poprawności klasyfikacji i czytelności wiedzy. Szczegóły w zadaniu na Moodle.