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 9

Temat
System Weka – wizualizacja granic klasyfikacji i wybieranie atrybutów

Wymagane wiadomości wstępne z wykładu
Problem klasyfikacji

Konfiguracja komputera
Podczas laboratorium wykorzystywany będzie system Weka.

Linki
http://www.cs.waikato.ac.nz/ml/weka/
http://archive.ics.uci.edu/ml/
Uwaga! Jeśli są problemy z dostępem do strony UCI można użyć proxy, np.
https://www.proxysite.com

Plan laboratorium

1. Uruchomić system Weka
2. **Wizualizacja granic klasyfikacji**
   - 2.1. Sprawdzić czy istnieje plik danych „iris.2D.arff". Jeśli nie, to stworzyć go, usuwając z pliku „iris.arff" dwa pierwsze atrybuty.
   - 2.2. Uruchomić moduł BoundaryVisualizer z menu Visualization modułu Weka GUI Chooser.
   - 2.3. Wybrać plik z danymi „iris.2D.arff".
   - 2.4. Wybrać klasyfikator weka.classifiers.rules.OneR.
     - 2.4.1. Zaznaczyć opcję Plot training data.
     - 2.4.2. Wcisnąć przycisk Start.
     - 2.4.3. Dlaczego wykres tak wygląda? Nauczoną regułę można sprawdzić w module Explorer.
     - 2.4.4. Powtórz powyższą procedurę dla IBk i k=1, 3, 5, 7.
     - 2.4.5. Dlaczego dla k=1 są tylko 3 kolory, a dla k>1 jest więcej odcieni? Co oznaczają dodatkowe obszary?
   - 2.5. Powtórz powyższą procedurę dla NaiveBayes.
     - 2.5.1. Ustaw useSupervisedDiscretization na true i ponownie wygeneruj diagram. Skąd różnica w wyglądzie?
     - 2.5.2. Dodaj dodatkowe przykłady dla klasy iris-versicolor klikając na rysunku i powtórz wizualizację.
   - 2.6. Wczytaj ponownie plik z danymi i narysuj diagram dla Jrip.
     - 2.6.1. Sprawdź w module Explorer jak wyglądają reguły.
     - 2.6.2. Spróbuj przekonwertować je do postaci bez ustalonego porządku.
   - 2.7. Powtórz procedurę dla J48.
     - 2.7.1. Wypróbuj różne wartości parametru minNumObj (3, 2, 1).
   - 2.8. Sprawdź jak zachowują się poszczególne algorytmy dla danych z dodanym szumem (por. 2.5.2).
3. **Wybieranie atrybutów**
   - 3.1. Uruchom moduł Explorer.
   - 3.2. Wczytaj zestaw danych labor.arff.
   - 3.3. Wybierz zakładkę Select attributes.
     - 3.3.1. Uruchom CfsSubsetEval. Celem jest znalezienie zestawu atrybutów silnie skorelowanych z kategorią, a słabo pomiędzy sobą.
     - 3.3.2. Zamień BestFirst na GreedyStepwise i ustaw searchBackwards na true. W ten sposób można uzyskać automatyzację procedury eliminacji wstecznej omawianej na poprzednim laboratorium.
     - 3.3.3. Uruchom InfoGainAttributeEval.
   - 3.4. Wczytaj zestaw danych diabetes.arff i sprawdź skuteczność klasyfikacji przy użyciu NaïveBayes (useSupervisedDiscretization=true) z użyciem kroswalidacji.
   - 3.5. Dodaj kopię pierwszego atrybutu (filtr weka.filters.unsupervised.attribute.Copy w zakładce Preprocess) i znowu skuteczność klasyfikacji przy użyciu NaïveBayes. Dodaj koleją kopię i znowu sprawdź.
   - 3.6. Wybierz zakładkę Classify
     - 3.6.1. Wybierz meta klasyfikator AttributeSelectedClassifer. Jego użycie gwarantuje dobór atrybutów z użyciem jedynie zbioru treningowego, bez podglądania testowego.
     - 3.6.2. W ustawieniach metaklasyfikatora wybierz klasyfikator NaïveBayes.
     - 3.6.3. Sprawdź działanie różnych metod poszukiwania. Dla Ranker ustaw numToSelect na 8 ponieważ tyle było oryginalnych atrybutów.
     - 3.6.4. Które metody usunęły kopie atrybutów?