WEKA
modelowanie bezrobocia z użyciem sztucznych sieci neuronowych (SSN)

1 WEKA — elementy potrzebne do zadania

W niniejszym rozdziale omówione zostaną pokrótce elementy systemu WEKA Explorer, które wykorzystane zostaną do badań nad danymi o bezrobociu.

1.1 Przygotowanie danych

Zakładka [`Preprocess`] pozwala na załadowanie danych z pliku, strony, bazy danych, wygenerowanie danych.

Grupa [`Current relation`] podaje nazwę wczytanego zbioru, liczbę atrybutów i liczbę rekordów.

Grupa [`Attributes`] pozwala na zarządzanie atrybutami, które są podane w liście. Można je dowolnie wybierać i usuwać.

Grupa [`Selected attribute`] podaje szczegółowe informacje dla wybranego atrybutu z listy w tym:

- nazwę
- typ
- liczba i procent brakujących danych (oznaczonych w pliku znakiem "?")
- liczbę różnych wartości dla atrybutu
- liczba i procent rekordów niepowtarzalnych, czyli takich, które posiadają taką wartość atrybutu, że inne rekordy takiej nie mają)
- inne dane: statystyczne (minimum, maksimum, średnia odchylenie standardowe) dla danych numerycznych i wartości i liczność rekordów z daną wartością dla danych nominalnych.

[`Histogram`] wyświetlany jest w dole okna w prawej jego części.

Przycisk [`Visualize all`] otwiera okno z histogramami dla wszystkich atrybutów.

Lista z wyborem atrybutu decyzyjnego (klasy) dla metod nadzorowanych znajduje się nad histogramem.

Zastosowanie filtrów [Apply] w sekcji [Filter] wpływa na zbiór danych w zależności od tego jaką metodę się zastosuję. Jest to między innymi zależne, czy wybrany jest atrybut klasy (decyzyjny).

1.2 Klasyfikator

Zakładka [Classify] pozwala na zastosowanie dla danych jednego z klasyfikatorów (wbudowanych i własnych).

Przycisk [Choose] służy do wyboru z drzewa jednej z metod. Jeżeli klasyfikator jest sparametryzowany, to w polu edycji pojawi się polecenie wywołania klasyfikatora z domyślną listą parametrów. Kliknięcie w owo pole otwiera dialog, gdzie można owe parametry zmieniać.

W grupie [Test] można dokonać testowania modelu na jeden z wybranych sposobów:

- testować na danych, na których odbyło się uczenie klasyfikatora
- podać niezależny plik z danymi do testowania
- użyć walidacji krzyżowej (http://pl.wikipedia.org/wiki/Sprawdzian_krzyÅijowy)
- dzieląc zbiór danych na grupę uczącą i testową określając ile procent przypada na dane uczące.

W liście należy wybrać zmienną atrybut wyjściowy (może być tylko jeden). Metody, które są zaimplementowane w WEKA mogą działać tylko dla zmiennych nominalnych i numerycznych lub dla obu.

Uczenie klasyfikatora rozpocznie się, gdy naciśnie się przycisk [Start].

Wyniki z uczenia i testowania modeli podawane są w oknie [The Classifier Output Text] i mogą zawierać:

- Run information: lista informacji o opcjach schematu uczenia modeli w tym: nazwa relacji, rekordów i trybu testowania.
- Classifier model (full training set): tekstowa reprezentacja modelu utworzonego wskutek uczenia.
- Summary: statystyki podsumowujące jak dobrze model działa na danych testowych.
- Detailed Accuracy By Class: dokładne statystyki rozdzielone na klasy.
- Confusion Matrix: pokazuje ile rekordów przypisano do każdej klasy. Właściwa klasa jest w wierszu, a wybrana przez model w kolumnie.

W [The Result List] widoczne są tworzone w trakcie pracy systemu modele. Kliknięcie prawym klawiszem myszki w tym obszarze powoduje dostęp do takich opcji jak: podgląd w głównym oknie, podgląd w oddzielnym oknie, zapisanie wyników, załadowanie nauczonego modelu, wykresy z błędem klasyfikacji (poprawna klasyfikacja - krzyżyk, niepoprawna kwadrat) i inne.

2 Dane o bezrobociu

W pliku bezrobicie.xls znajdują się dane o bezrobociu w Polsce na przełomie lat 1992-2009. Każdy rekord, to dane z miesiąca w roku. Plik zawiera 216 próbek z 26 atrybutami (**od X1 do X26**) i zmienną decyzyjną (**Y**) „stopa bezrobocia".

Dane są znormalizowane (przyjmują wartości z przedziału od 0 do 1). Zmienna decyzyjna (wyjście) jest typu „numeric".

Dane stanowią zbiór uczący/testujący dla sztucznej sieci neuronowej. Należy odpowiednio go przygotować przystosowując do formatu akceptowanego przez WEKA. Należy przygotować oba formaty:

1. csv: zapisać arkusz do formatu csv pozostawiając wiersz z opisem atrybutów
2. arff: z pliku csv stworzyć plik w formacie arff [`WEKA: funkcja Save...`], który jest szczegółowo opisany pod adresem: `http://www.cs.waikato.ac.nz/~ml/weka/arff.html`

3 Zadania do wykonania

Głównym celem badań jest zamodelowanie stopy bezrobocia sztuczną siecią neuronową (zadanie regresji). Innymi słowy SSN ma znaleźć zależność danych wejściowych (atrybutów) od wyjścia. Celem pośrednim jest zbadanie wrażliwości modelu na architekturę i parametry uczenia.

Należy pamiętać, by uczciwie ocenić wyniki powinno się uruchomić każdy z wariantów chociaż dziesięciokrotnie z różną wartością [`seed`] (parametr klasyfikatora). Takie zmiany spowodują zróżnicowanie w kolejności danych uczących i testowych oraz w wartościach początkowych wag, co spowoduje różne poziomy dopasowania modelu. Można wykorzystać moduł [`Experimenter`], który pozwala na wielokrotne przeprowadzanie badań i testowanie kilku modeli oraz zapisuje wyniki wraz ze statystyki i średnim dopasowaniem każdego modelu (można też uzyskać informacje, czy różnice pomiędzy modelami są istotne statystycznie.

3.1 Zadanie 1 — Badanie modelu MultilayerPerceptron

Korzystając z dokumentacji/innych źródeł krótko opisać w sprawozdaniu jakie cechy ma stosowana w WEKA sieć MLP.

1. Pozwolić zbudować automatyczny (autoBuild=True) model wyświetlając jego graficzną postać (GUI true). Wizualizacja ułatwia zrozumienie działania modelu i pozwala na ręczną jego modyfikację.
2. Zamiana architektury sieci: ustalanie różnej liczby neuronów w warstwach ukrytych (w sprawozdaniu zawrzeć wyjaśnienie dla opcji): „a", „i", „o", „t", oraz dwa własne warianty z jedną i dwoma warstwami ukrytymi (wybraną liczbę neuronów uzasadnić wynikami z poprzednich eksperymentów). W sprawozdaniu zawrzeć wyniki z badań nad tymi parametrami.

3. Wpływ współczynnika uczenia na model. Zmiana parametru z wybranym zakresie (0-1). Ustawić współczynnik uczenia na 1 i pozwolić mu się automatycznie zmniejszać [`decay - True`].

4. Wpływ współczynnika momentum na model. Zmiana parametru z zakresie (0-1).

3.2 Zadanie 2 — Regresja z użyciem RBFNetwork

Korzystając z dokumentacji/innych źródeł krótko opisać w sprawozdaniu jakie cechy ma stosowana w WEKA sieć RBF.

1. Badanie wpływu minimalnego odchylenia standardowego dla klastrów na modelowanie poprzez zmianę jego warości.

2. Zamiana architektury sieci i wpływ na wyniki modelowania. Ustalanie liczby klastrów (1-liczba wejść).

3.3 Sprawozdanie

Sprawozdanie w formacie i o nazwie *imie_ nazwisko.pdf* należy przesłać w terminie do 10-go listopada na adres jkolodziejczy[at]wi.zut.edu.pl. Opóźnienia będą wpływały na obniżenie punktacji za sprawozdanie.

Każde badanie powinno być krótko opisane (wartości parametrów, liczba prób, itd.) i prezentować przebieg prób modelowania w postaci czytelnej tabeli i/lub wykresu oraz zawierać interpretację wyników (wnioski). Postarać się zauważyć jakieś tendencje, wyznaczyć zestaw parametrów najkorzystniejszych.

Wszelkie plagiaty oceniane będą na 0 punktów (niezależnie od autora).

4 Pytania na wejściówkę

1. Na czym polega walidacja krzyżowa?

2. Z jakiego powodu i jak przeprowadza się normalizację danych przed ich modelowaniem?

3. Jaka jest różnica pomiędzy formatem arff i sparse arff?

4. Jakie znaczenie ma współczynnik uczenia, jakie wartości tego współczynnika stosuje się najczęściej i dlaczego?

5. Jak momentum wpływa na uczenie modelu i jakie wartości tego współczynnika stosuje się najczęściej i dlaczego?

6. Jaka jest różnica pomiędzy: Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error?

7. Najważniejsze cechy (architektura, metody uczenia) MLP i możliwe ich zastosowania

8. Najważniejsze cechy (architektura, metody uczenia) RBF i możliwe ich zastosowania.