WEKA
klasyfikacja z użyciem sztucznych sieci neuronowych

1 WEKA — elementy potrzebne do zadania

WEKA (Data mining software in Java `http://www.cs.waikato.ac.nz/ml/weka/`) jest narzędziem zawierającym zbiór narzędzi do wykonywania zadań drążenia danych (data mining). Program należy pobrać i zainstalować na swoim dysku. Uruchomić WEKA Explorer.

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

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 uczące i testujące

W zadaniu wykorzystane zostaną dane medyczne, które zawierają dane o pacjencie (dane ogólne) i zarejestrowane podczas wizyty u lekarza symptomy oraz wyniki testów np. pomiar ciśnienia, czy wyniki diagnostyczne.

Wszystkie dane z baz medycznych muszą być przekonwertowane na dane numeryczne. Bardzo często wartości atrybutów opisujących symptomy są binarne (1— symptom wystąpił, 0 - nie wystąpił).

Korzystając z `http://archive.ics.uci.edu/ml/` UCI Repository of Machine Learning Databases pobrać zbiór danych: Wisconsin Diagnostic Breast Cancer (WDBC) (Original) z dziesięcioma atrybutami (breast-cancer-wisconsin.data).

Dane stanowią zbiór uczący/testujący dla sztucznej sieci neuronowej. Należy odpowiednio go przygotować przystosowując do formatu akceptowanego przez WEKA. Plik WDBC otworzyć w dowolnym arkuszu kalkulacyjnym i zapisać do formatu csv dodając uprzednio wiersz z opisem atrybutów (opis atrybutu na stronie www lub w breast-cancer-wisconsin.names)

W sprawozdaniu należy opisać zbiór danych w następujący sposób:

1. Jakiego problemu klasyfikacji dotyczą dane (jakie są atrybuty i jakie są klasy)?
2. Jakiego typu są poszczególne atrybuty?
3. Czy w pliku są brakujące wartości? Ile procent?
4. Jaka jest liczność zbioru próbek?
5. Ile jest próbek w poszczególnych klasach?
6. Czy można z góry określić, że w zbiorze są atrybuty nie wpływające na klasyfikację?

3 Zadania do wykonania

Głównym celem badań jest zastosowanie wielowarstwowego perceptronu do zadania klasyfikacji pacjentek z rakiem piersi. Pośrednim celem jest zbadanie wrażliwości modelu na różne parametry: architektura sieci, współczynnik uczenia, współczynnik momentum.

Należy pamiętać, by uczciwie ocenić wyniki powinno się uruchomić każdy z wariantów chociaż kilkukrotnie 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. Do wielokrotnych testów wygodniej 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 Szczegółowy opis zadania

1. Wykonywać pięciokrotną kroswalidację. Pozwolić zbudować automatyczny (autoBuild=True) model wyświetlając jego graficzną postać (GUI true). Wizualizacja ułatwia

zrozumienie działania modelu i pozwala na ręczną jego modyfikację. W ramach tego zadania wykonać wybór najlepszych parametrów uczenia i architekturę SSN:

(a) Zamienić architekturę sieci poprzez zmienianie liczby warstw i liczby neuronów w warstwach ukrytych. Przetestować opcje „a", „i", „o", „t", oraz wartości „0" i "2,2" . W sprawozdaniu zawrzeć opis opcji i wyniki z badań nad tymi parametrami. Skomentować rezultaty.

(b) Wpływ współczynnika uczenia na model badać zmieniając parametr *learningRate* w zakresie od zera do jeden z wybranym krokiem. Ustawić współczynnik uczenia na 0.9 i pozwolić mu się automatycznie zmniejszać [`decay - True`].

(c) Wpływ współczynnika momentum na model badać zmieniając parametr w zakresie od zera do jeden z wybranym krokiem.

3.2 Sprawozdanie

Sprawozdanie w formacie i o nazwie *imie_ nazwisko.pdf* należy przesłać na adres jkolodziejczyk[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.

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

4 Pytania na wejściówkę

1. Na czym polega walidacja krzyżowa?
2. Jakie stosuje się sposoby testowania klasyfikatora (w tym walidacja krzyżowa)?
3. Z jakiego powodu i jak przeprowadza się normalizację danych przed ich modelowaniem?
4. Co to jest i na czym polega wybór atrybutów?
5. Jaka jest różnica pomiędzy formatem arff i sparse arff?
6. Jakie znaczenie ma współczynnik uczenia, jakie wartości tego współczynnika stosuje się najczęściej i dlaczego?
7. Jak momentum wpływa na uczenie modelu i jakie wartości tego współczynnika stosuje się najczęściej i dlaczego?
8. Co to jest architektura sieci neuronowej?
9. Co opisuje Confusion matrix?
10. Co to jest histogram dla zmiennej?
11. Co to jest dyskretyzacja?

12. Jak jest obliczane True Positive Rate i jaka jest jego interpretacja?
13. Jak jest obliczane False Positive Rate i jaka jest jego interpretacja?
14. Jak jest obliczane Precision i jaka jest jego interpretacja?
15. Jak jest obliczane Recall i jaka jest jego interpretacja?
16. Jakie miary wykorzystuje się przy ocenie klasyfikatora?