

Ocena modeli¹

I. Ocena modeli

Bardzo ważne jest pytanie, jak bardzo możemy ufać otrzymanym klasyfikatorom w stosunku do przyszłych danych. Oczywiście przyszłych danych nie znamy. Możemy jednak próbować estymować parametry mówiące o jakości klasyfikacji przyszłych przykładów. Służy do tego stosunkowo prosty pomysł: podział zbioru uczącego na część trenującą i testującą. Na tym pomysle oparta jest walidacja skrośna.

1. Walidacja skrośna dla zbioru hepatitis.arff.

1. Na przestrzeń roboczą wstaw węzeł *Workflow Control*→*Meta Nodes*→*Loop x-times*. Zapoznaj się z opisem tego węzła. Zwróć uwagę na fakt, że jest to w rzeczywistości grupa węzłów (wirtualny pod-przepływ). Węzeł ten służy do wykonywania pewnych działań wskazaną liczbę razy.
2. Połącz wejście węzła *Loop x-times* do wyjścia węzła *ARFF Reader* (skonfigurowanego do odczytu pliku *hepatitis.arff*)
3. "Wejdź" do węzła *Loop x-times* (dwukrotne kliknięcie)
4. Zapoznaj się z opisami węzłów *Counting Loop Start* i *Loop End*. Zajrzyj do konfiguracji węzła *Counting Loop Start* i upewnij się, że liczba powtórzeń ustawiona jest na 10.
5. Na przestrzeń roboczą (pod-przepływu) przesunij węzeł *Analytics*→*Mining*→*Scoring*→*Meta Nodes* →*Cross Validation*. Zapoznaj się z opisem węzła *Cross Validation*. Zwróć uwagę na fakt, że jest to w rzeczywistości grupa węzłów, służąca do przeprowadzenia walidacji skrośnej.
6. Wejście węzła *Cross Validation* połącz z wyjściem *Counting Loop Start*, zaś "dolne" wyjście z wejściem węzła *Loop End*.
7. "Wejdź" do węzła *Cross Validation*.
8. Zapoznaj się z opisami węzłów *X-Partitioner* i *X-Aggregator*. Pierwszy z węzłów dokonuje podziału (lub podziałów) zbioru uczącego na część trenującą i testującą. Drugi służy do zbierania wyników klasyfikacji.
9. Pozostaw w przepływie utworzone domyślnie węzły *Decision Tree Learner* i *Decision Tree Predictor*.
10. Zapoznaj się z opcjami konfiguracyjnymi węzła *X-Partitioner*. Upewnij się, czy liczba walidacji wynosi 10. Zwróć uwagę, że domyślnie wybierane jest próbkowanie bez stratyfikacji. Zmień to ustawienie, wybierając opcję *Stratified sampling*.
11. Skonfiguruj węzeł *Decision Tree Learner*. Upewnij się, czy jako atrybut celu wybrany został *Class*. Pozostałe opcje pozostaw domyślne.
12. Skonfiguruj węzeł *X-Aggregator*. Jako atrybut celu podaj atrybut *Class*, jako atrybut predykcyjny *Prediction (Class)*.
13. Wróć do głównego przepływu. Do wyjścia węzła *Loop x-times* podłącz nowe węzły : *Interactive Table (local)* oraz *Statistics*.

¹ Przykłady zostały opracowane na podstawie: Kursu IBM: Introduction to IBM SPSS Modeler and Data Mining (Student Guide) oraz książki Daniela T. Larose „Odkrywanie wiedzy z danych” Wprowadzenie do eksploracji danych. Wydawnictwo Naukowe PWN, Warszawa 2006.

14. Uruchom węzeł *Loop x-times* (dla lepszych wrażeń wizualnych można w trakcie pracy obejrzeć jego pod-przepływ).
15. Obejrzyj węzeł *Interactive table (Execute and Open Views)*.
16. Obejrzyj węzeł *Statistics*.
17. Zresetuj węzeł *Loop x-times*. Ponownie uruchom i obejrzyj węzeł *Statistics*. Powtórz proces kilka razy.
18. Dokonaj oceny modelu *Decision Tree Learner* z ustawionym przycinaniem na MDL.
19. Zresetuj węzeł *Loop x-times*. Ponownie uruchom i obejrzyj węzeł *Statistics*. Powtórz proces kilka razy.
20. Przebuduj klasyfikator wewnątrz *Cross Validation*, wykorzystując Naiwny Klasyfikator Bayesa. Oceń go.

2. Ocena modeli dla klasyfikatorów złożonych.

Klasyfikatory złożone to klasyfikatory składające się z kilku klasyfikatorów. Budując takie konstrukcje, liczymy na to, że właściwości połączonego zestawu klasyfikatorów sprawiają, że uzyskamy lepsze własności predykcyjne. Stosuje się różne metody łączenia klasyfikatorów; tutaj wykorzystamy *bagging* i *boosting*.

1. Na przestrzeń roboczą dołóż węzeł *Analytics*→*Integrations*→*Weka*→*Classification algorithms*→*meta*→*Bagging*. Połącz go z węzłem ARFF Reader.
2. Zapoznaj się opisem węzła *Bagging*. Wejdź do opcji konfiguracyjnych węzła. Jako klasyfikator wybierz *weka*→*classifiers*→*trees*→*J48*. Jako *bagPercentSize* (liczbę przykładów w jednym zestawie) podaj 40. Upewnij się, że jako atrybut celu wybrany został *Class*.
3. Uruchom węzeł *Bagging* i zapoznaj się z wynikami.
4. Na przestrzeń roboczą pod-przepływu dołóż węzeł *Analytics*→*Integrations*→*Weka*→*Classification algorithms*→*meta*→*AdaBoostM1*. Połącz go z węzłem ARFF Reader.
5. Zapoznaj się opisem węzła *AdaBoostM1*. Wejdź do opcji konfiguracyjnych węzła. Jako klasyfikator wybierz *weka*→*classifiers*→*trees*→*J48*. Upewnij się, że jako atrybut celu wybrany został *Class*.
6. Uruchom węzeł *AdaBoostM1* i zapoznaj się z wynikami.

3. Zadanie samodzielne

Zbuduj trzy zestawy węzłów służące do oceny (za pomocą walidacji skrośnej) odpowiednio: klasyfikatora J48, klasyfikatora złożonego wykorzystującego *bagging* i klasyfikatora złożonego wykorzystującego *boosting* (działających jak te skonfigurowane w pkt. 2).