Reguły asocjacyjne i poprawianie jakości klasyfikatorów

Niniejsze laboratorium poświęcimy przećwiczeniu kilku ostatnich opisanych na wykładach mechanizmów: regułom asocjacyjnym i doborowi atrybutów oraz parametrów klasyfikacji.

Reguły asocjacyjne

Reguły asocjacyjne służą odnajdywaniu zależności pomiędzy wartościami poszczególnych atrybutów bez wskazywania konkretnego atrybutu celu. Reguły asocjacyjne są przydatne dla analityka (lub jego klienta), ze względu na to, że pozwalają na określenie pewnych charakterystycznych prawidłowości zachodzących w zbiorze przykładów. Reguły asocjacji są także często stosowane jako narzędzie wspierające EDA (eksploracyjną analizę danych).

1. Po uruchomieniu systemu przenieś na przestrzeń roboczą węzeł *ARFF Reader*. Skonfiguruj go do odczytu pliku *Hepatitis.arff* (jak pamiętasz, zawiera on dane na temat pacjentów chorych na żółtaczkę i wyników ich terapii).
2. Będziemy wykorzystywać algorytm Apriori, zatem ze zbioru należy usunąć atrybuty numeryczne (nienominalne). Przesuń na przestrzeń roboczą węzeł *Manipulation* → *Column* → *Filter* → *Column Filter*.
3. Na wejście węzła *Column Filter* podaj wyjście węzła *ARFF Reader*. Skonfiguruj węzeł *Column Filter* tak, aby usuwał atrybuty numeryczne (z ikonką I lub D). Uruchom węzeł *Column Filter*.
4. Na przestrzeń roboczą dołóż węzeł *Analytics* → *Mining* → *Weka* → *Association Rules* → *Apriori*. Na wejście węzła podaj wyjście węzła *Column Filter*.
5. Zapoznaj się z opcjami konfiguracyjnymi węzła *Apriori*. Upewnij się, czy liczba reguł wynosi 10.
6. Uruchom węzeł *Apriori* poprzez opcję *Execute and Open Views*.
7. Zapoznaj się z otrzymanymi wynikami, zwracając uwagę na to, jakie reguły zostały wygenerowane, jakie jest minimalne pokrycie i jaka jest poprawność reguł.
8. Skonfiguruj węzeł *Apriori*, zmieniając liczbę reguł na 100.
9. Uruchom węzeł *Apriori* poprzez opcję *Execute and Open Views*.
10. Ponownie zapoznaj się z otrzymanymi wynikami.

Dobór parametrów procesu klasyfikacji

Fakt, że dysponujemy narzędziem oceny jakości klasyfikatorów, otwiera nam drogę do zastosowania wielu ciekawych automatycznych technik doboru parametrów procesu klasyfikacji. Droga ta jest koncepcyjnie bardzo prosta: dla różnych zestawów parametrów budujemy klasyfikatory i oceniamy, który jest najlepszy. Tutaj przedstawimy dwa sposoby zastosowania tej techniki: do doboru parametrów pracy algorytmu oraz (nieco później) do doboru atrybutów metodą "wstecz".

Dobór parametrów pracy

1. Na przestrzeń roboczą wstaw węzeł *Analytics* → *Mining* → *Weka* → *Classification Algorithms* → *Meta* → *CVParameterSelection*. Na jego wejście podaj wyjście węzła *ARFF Reader*.
2. Zapoznaj się z opcjami konfiguracyjnymi węzła *CVParameterSelection*. Jako klasyfikator wybierz *weka* → *classifier* → *trees* → *J48*. Zauważ, że przy nazwie algorytmu pojawiły się domyśle wartości parametrów (symbol C: *Confidence factor* oraz symbol M: *Min num obj*).
3. Skonfiguruj listę *CVParametres* (nadal w opcjach konfiguracyjnych *CVParameterSelection*). Dodaj do listy ciąg znaków "C 0.15 0.35 10". Oznacza to, że będziemy szukać najlepszej wartości parametru C (*Confidence factor*) w przedziale 0.15–0.35, używając 10 iteracji.
4. Uruchom węzeł za pomocą opcji *Execute and Open Views*. Zwróć uwagę na to, jaka wartość parametru została dobrana jako optymalna.

Dobór atrybutów metoda „wstecz"

1. Przenieś na przestrzeń roboczą węzeł *Analytics* → *Mining* → *Feature Selection* → *Meta Nodes* → *Backward Feature Elimination*. Na oba wejścia węzła podaj dane z węzła *ARFF Reader*.
2. Zapoznaj się opisem węzła *Backward Feature elimination*. "Wejdź" do węzła i zapoznaj się z opisami węzłów wewnątrz metawęzła.
3. Skonfiguruj węzeł *Feature Selection Loop Start*, tak aby włączone do procesu były wszystkie atrybuty z wyjątkiem atrybutu celu *Class*.
4. Skonfiguruj węzeł *Naive Bayes Learner*, upewniając się, czy jako atrybut celu wybrany został *Class*. Zaznacz też opcję *Ignore missing values*.
5. Skonfiguruj węzeł *Scorer* tak, aby za atrybut celu (*First column*) uznawał on *Class*, a za atrybut predykcyjny (*Second column*) *Prediction (Class)*.

6. Skonfiguruj węzeł *Feature Selection Loop End*, tak aby minimalizował on błąd klasyfikacji (wybierz w polu *Score* opcję *Error* i zaznacz pole *Minimize score*).
7. Uruchom węzeł *Feature Selection Filter*. Poczekaj na zakończenie procesu doboru atrybutów.
8. Skonfiguruj węzeł *Feature Selection Filter*. Zobacz, jaki wzrost błędu powoduje ujmowanie kolejnych atrybutów. Zwróć uwagę na to, który zestaw atrybutów jest najlepszy. Wybierz ten zestaw atrybutów i zatwierdź, klikając *OK*.
9. Uruchom ponownie węzeł *Feature Selection Filter*. Obejrzyj odfiltrowane dane, wybierając z menu kontekstowego węzła *Feature Selection Filter* opcję *Filtered table*.