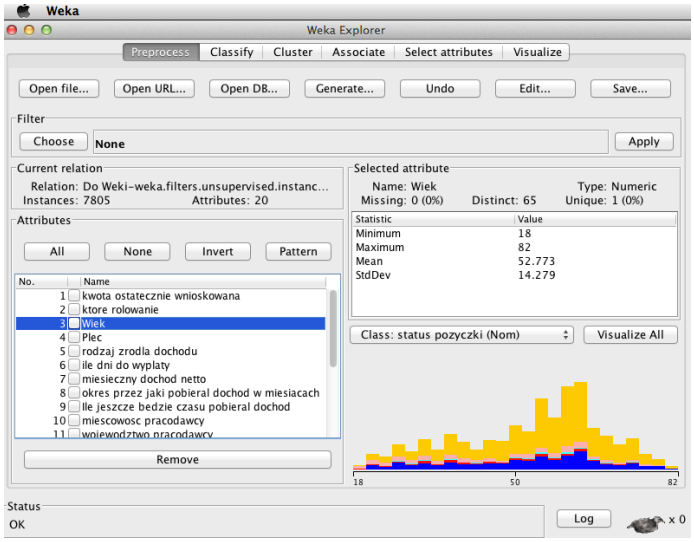
Lista 3 - Przetwarzanie danych

Przetwarzanie danych z wykorzystaniem GUI Weki

Należy otworzyć skonstruowany w ramach poprzedniej listy zbiór danych *XXXXXXL2_2.arff* poprzez udostępnione przez *Wekę* GUI. W ramach zbioru danych możliwe jest przeglądanie statystyk dotyczących poszczególnych atrybutów. W przypadku atrybutów rzeczywistych wyświetlanie są statystyki dotyczące średniej, minimalnej, maksymalnej wartości atrybutu, oraz odchylenia standardowego. W przypadku atrybutów nominalnych podawane są częstości pojawiania się zadanych wartości. Ostatni atrybut jest domyślnie traktowany jako atrybut klasy (wyjścia, wyniku podejmowania decyzji, rezultatu klasyfikacji, bądź regresji).



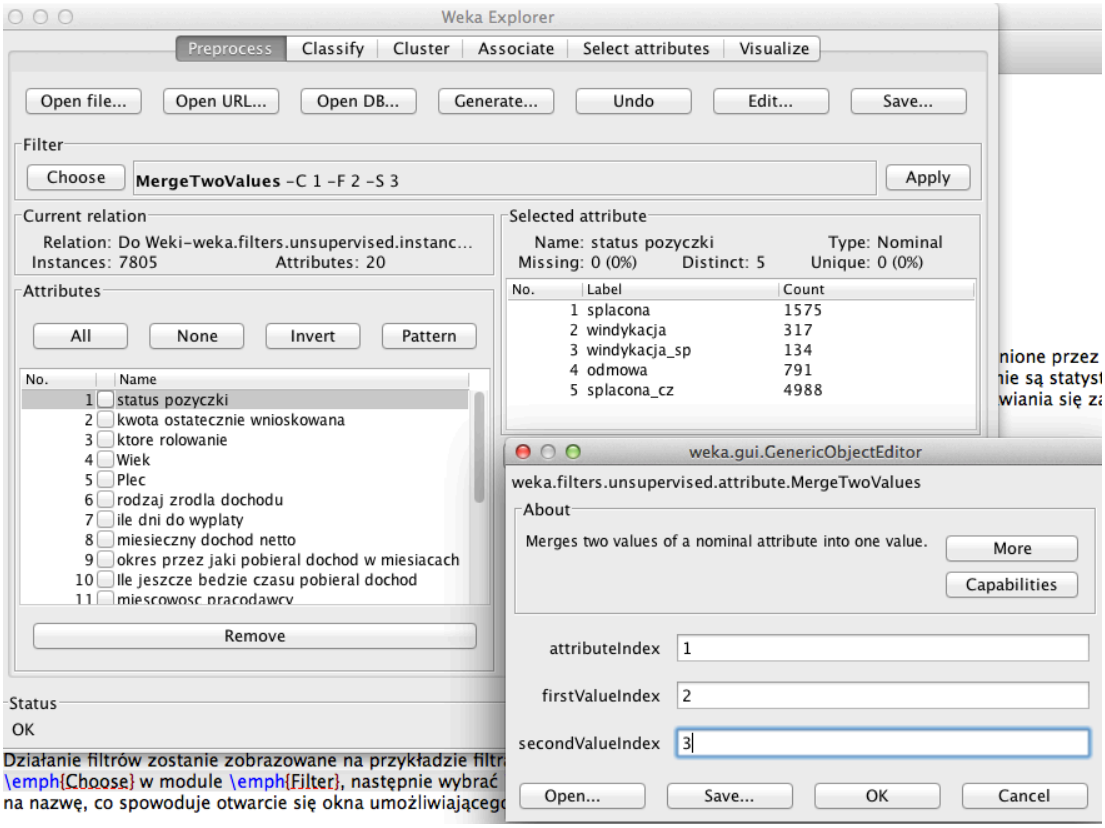
Kluczowym modułem wykorzystywanym do przetwarzania danych jest moduł *Filter*. W ramach tego modułu dostępne są filtry pozwalające m. in. na próbkowanie danych, eliminacji brakujących wartości, generowanie syntetycznych danych, dodawanie i usuwanie atrybutów, dyskretyzację i binaryzację atrybutów numerycznych, łączenie dwóch wartości atrybutów nominalnych, oraz wiele innych operacji na danych, oraz atrybutach. Filtry są w *Wece* kategoryzowane jako nadzorowane (*ang. supervised*), oraz nienadzorowane (*ang. unsupervised*). Pierwszą grupę stanowią filtry, które wykorzystują wiedzę na temat wyjścia (klasy), drugą grupę metod stanowią natomiast filtry które działają bez wiedzy na temat klasy. Filtry mogą dotyczyć zarówno atrybutów (ich usuwania, dodawania, łączenia ich wartości nominalnych), jak i samych danych (próbkowanie, normalizacja).

Przykład wykorzystania filtra

Działanie filtrów zostanie zobrazowane na przykładzie filtra *MergeTwoValues*, służącego do łączenia dwóch wartości atrybutu nominalnego. Należy wybrać opcję *Choose* w module *Filter*, następnie wybrać *unsupervised-attribute-MergeTwoValues*. W oknie obok przycisku pojawi się nazwa wybranego filtra. Należy kliknąć na nazwę, co spowoduje otwarcie okna umożliwiającego zdefiniowanie parametrów wybranego *Filtra*.

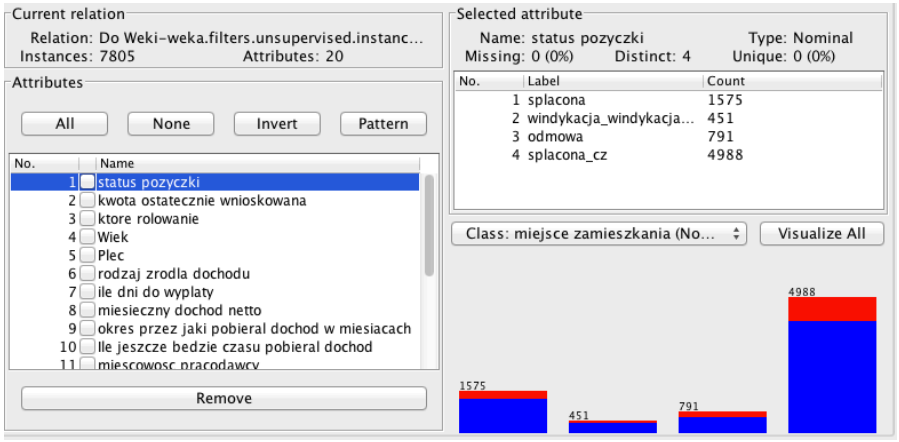
Filtr *MergeTwoValues* wymaga zdefiniowania trzech parametrów:

- Indeks atrybutu (*attributeIndex*), którego wartości będą łączone (indeksowanie atrybutu następuje od 1).



- Indeks pierwszej wartości atrybutu nominalnego (*firstValueIndex*), który będzie podlegać złączeniu.
- Indeks drugiej wartości atrybutu nominalnego (*secondValueIndex*), który będzie podlegać złączeniu.

Po ustaleniu wartości parametrów należy wcisnąć przycisk *Apply*. Po wciśnięciu przycisku, w rozpatrywanym przykładzie nastąpi złączenie 2 i 3 wartości (**windykacja**, oraz **windykacja_sp**) atrybutu o indeksie 1 (**status pożyczki**). W rezultacie, wszystkie instancje zbioru danych, dla których atrybut **status pożyczki** przyjmował wartości **windykacja**, bądź **windykacja_sp** zostaną zastąpione wartością **windykacja_windykacja_sp**.



Analogicznie wykorzystuje się inne filtry dostępne w środowisku *Weka*. Aby dowiedzieć się więcej na temat danego filtra należy wcisnąć przycisk *More* w oknie specyfikacji wartości parametrów danego filtra. Przetwarzany zbiór danych można w każdej chwili zapisać wciskając przycisk *Save...*

Przetwarzanie danych z wykorzystaniem Weki, jako biblioteki programistycznej

Narzędzie *Weka* może zostać również wykorzystane jako biblioteka programistyczna w języku *Java*. Pozwala to na sprawne łączenie i konstrukcję nowych funkcjonalności związanych z analizą danych. Aby wykorzystać narzędzie jako bibliotekę programistyczną należy utworzyć projekt w dowolnym środowisku umożliwiającym programowanie w języku *Java* (*NtBeans*, *Eclipse*), i zaimportować do projektu plik *weka.jar*. Podstawowe funkcje związane z wykorzystaniem *Weki*, jako biblioteki programistycznej zostały opisane pod adresem `http://weka.wikispaces.com/Use+WEKA+in+your+Java+code`.

Zadania

Przetwarzanie danych z wykorzystaniem GUI Weki

Należy otworzyć skonstruowany w ramach poprzedniej listy zbiór danych *XXXXXXL2_2.arff* poprzez udostępnione przez *Wekę* GUI. Na wczytanym zbiorze danych należy wykonać następujące operacje:

1. Zmniejszyć liczbę wartości atrybutu **status pożyczki** do dwóch (*2 pkt*):
   - **dobry**, gdy atrybut przyjmował wartości **splacona**, **splacona_cz**;
   - **zly**, gdy atrybut przyjmował pozostałe wartości;
2. Należy usunąć atrybut **opóźnienie spłaty**. Następnie należy zmienić kolejność atrybutów tak, by atrybut **status pożyczki** znalazł się na ostatnim miejscu (Należy wykorzystać filtr *Reorder*) - *1 pkt.*
3. Należy Dokonać dyskretyzacji atrybutu **Wiek** na 5 równych przedziałów (Należy wykorzystać filtr *Discretize*) - *1 pkt.*

Należy zapisać przetworzony zbiór danych jako *XXXXXXL3_1.arff*.

Przetwarzanie danych z wykorzystaniem Weki, jako biblioteki programistycznej

1. Należy napisać skrypt w języku *Java*, który wczyta zbiór danych z pliku *XXXXXXL2_2.arff*, oraz przetworzy go w następujący sposób (*4 pkt*):
   - usunie rekordy, dla których wartość atrybutu **status pożyczki** wynosi **odmowa**, i te rekordy, dla których wartość pożyczki jest wyższa niż 900 zł.
   - usunie atrybut **status pożyczki**.
   - Zapisze zbiór danych jako *XXXXXXL3_2.arff*.