Podstawowe wstępne przetwarzanie danych : RapidMiner i Orange Data Mining

W pierwszym tygodniu zapoznamy się z dwoma narzędziami przydatnymi w pracy każdego górnika danych, będą to Orange Data Mining i RapidMiner. W drugiej części laboratorium zobaczymy, jak podstawowe operacje na danych można przeprowadzić w języku Python.

Orange 3

- Uruchom program Orange
- Umieść w przepływie operator File i otwórz w nim plik heart disease dostarczony z programem (wyjaśnienie znaczenia atrybutów [http://www.cs.auckland.ac.nz/courses/compsci367s2c/assignments/Pat.d/2012.d/ML/heart-c.arff])
- Prześlij wynik operatora File do operatora Data Table i zapoznaj się z charakterystyką zbioru danych
- Umieść w przepływie operator Box Plot i wyślij tam dane. Porównaj sposób wyświetlania zmiennych numerycznych i kategorycznych, bez i z włączonym podziałem na podgrupy wg atrybutu *płeć*
- Użyj operatora Distributions do porównania rozkładu wieku kobiet i mężczyzn
- Użyj operatora Scatter Plot aby wyświetlić związek między wiekiem i tętnem spoczynkowym dla kobiet i mężczyzn. Postaraj się także nanieść na wykres informacje o poziomie cholesterolu.
- Dodaj operator Discretize i jego wynik prześlij do operatorów Data Table i Distributions.

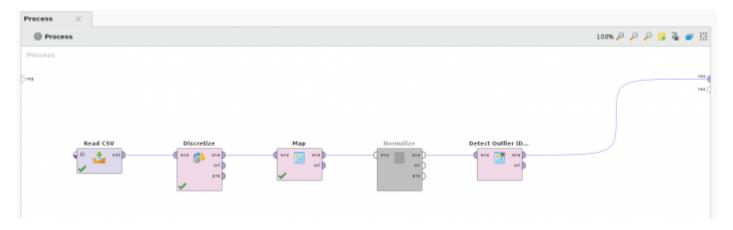
 Porównaj różne sposoby dyskretyzacji atrybutu wiek
- Użyj operatora Continuize aby obejrzeć wynik normalizacji atrybutu, zastanów się, które atrybuty powinny być normalizowane do przedziału [0-1], a które powinny podlegać standaryzacji

RapidMiner

- Uruchom program RapidMiner i wybierz opcję utworzenia nowego procesu.
- Zaimportuj do narzędzia dane o stosowaniu środków antykoncepcyjnych w Indonezji (tutaj znajdziesz wyjaśnienie atrybutów [http://b-course.hiit.fi/obc/cmcexpl.html])
- Prześlij wynik operatora Read CSV do prawej krawędzi przepływu i uruchom przepływ. Zapoznaj się z charakterystyką zbioru danych. Dokonaj wizualizacji danych.
- Powróć do widoku projektu procesu. Przejdź do zakładki *Operators*. W polu filtru wpisz łańcuch znaków *normal*. Lista dostępnych operatorów ograniczy się do dwóch pozycji. Przeciągnij operator Normalize i upuść go na przepływ danych z operatora Read CSV do rezultatu. Alternatywnie, możesz upuścić operator Normalize gdziekolwiek w panelu edycji procesu, a następnie przeciągnąć przepływ danych z portu wyjściowego (*out*) operatora Read CSV do portu wejściowego (*exa*) operatora Normalize. W tym drugim przypadku pamiętaj, aby port wyjściowy *exa* operatora Normalize połączyć z portem wynikowym *res*.
- Zaznacz operator Normalize. Wskaż, że chcesz normalizować jedynie atrybuty numeryczne (attribute filter type = value_type, value type = numeric). Jako metodę normalizacji pozostaw Z-transformation. Uruchom proces i zaobserwuj wynik. Czy potrafisz zgadnąć, co się stało?
- Zmień rodzaj transformacji na transformację zakresową do zakresu <0,1>. Porównaj otrzymany wynik z transformacją proporcjonalną. Wyłącz operator wybierając z menu kontekstowego opcję Enable operator lub korzystając ze skrótu klawiszowego Ctrl+E.
- Wróc do zakładki Operators i wyszukaj operatorów o nazwie Discretize.... Najpierw użyj operatora Discretize by Binning aby podzielić wiek kobiet na trzy przedziały. Następnie dodaj operator Mapiza pomocą pola value mappings dokonaj przetłumaczenia nazw zakresów wieku na wartości opisowe (np. młode, średnie, starsze)
- Wyłącz operator Discretize by Binning i w jego miejsce wstaw operator Discretize by Frequency, również wskazując trzy przedziały dyskretyzacji dla atrybutu age. Alternatywnie, możesz kliknąć

prawym klawiszem myszy na operatorze Discretize by Binning i z menu kontekstowego wybrać opcję Replace Operator, nawigując kolejno do Data Transformation/Type Conversion/Discretization/Discretize by Frequency.

- Wyszukaj operator Detect Outliers (Distance) i dodaj go do procesu. Wskaż, że detekcja wartości odstających odbywa się przez policzenie odległości do trzech najbliższych sąsiadów, oraz że w zbiorze danych występują trzy wartości osobliwe. Uruchom proces i zaobserwuj wynik. Przejdź do widoku wykresu i wybierz jako typ wykresu Scatter 3D Color. Postaraj się znaleźć taką kombinację atrybutów, które przekonująco wskazują, że znalezione lobiety faktycznie odstają od reszty.
- Dodaj do procesu operator Nominal to Binominal i wskaż atrybut standard jako atrybut do transformacji.
- Ostatecznie Twój przepływ powinien wyglądać następująco:



laboratorium1.txt · Last modified: 2020/03/19 13:05 by Mikołaj Morzy

Except where otherwise noted, content on this wiki is licensed under the following license:Public Domain [http://creativecommons.org/licenses/publicdomain/]