Segmentacja

Segmentacja (klasteryzacja) służy podziałowi zbioru uczącego na kilka kategorii bez wskazania atrybutu celu. Wyniki klasteryzacji mogą być bardzo przydatne dla analityka (lub jego klienta), jako że pozwalają wykryć charakterystyczne grupy przykładów (np. pacjentów, kredytobiorców, dostawy itp.). Generalną zasadą stosowaną przy algorytmach tworzenia modeli grupujących jest dążenie do jak najmniejszych różnic w ramach grupy przykładów oraz maksymalizowanie różnic pomiędzy grupami.

W ramach niniejszego ćwiczenia wykorzystamy algorytmy E-M i K-średnich, a wcześniej przygotujemy dane do klasteryzacji.

1. Przygotowanie danych

1. Po uruchomieniu systemu i wybraniu odpowiedniej perspektywy (KNIME) przenieś na przestrzeń roboczą węzeł *File Reader*. Skonfiguruj go do odczytu pliku *churn.txt* (jak pamiętasz, zawiera on dane na temat klientów firmy telekomunikacyjnej i ich rezygnacji z usług).
2. Będziemy wykorzystywać algorytmy wymagające atrybutów numerycznych, zatem ze zbioru należy usunąć atrybuty tekstowe (nienominalne). Zajmiemy się tym w kolejnych krokach.
3. Przesuń na przestrzeń roboczą węzeł *Manipulation → Column → Filter → Column Filter*.
4. Na wejście węzła *Column Filter* podaj wyjście węzła *File Reader*. Skonfiguruj węzeł *Column Filter* tak, aby usuwał atrybuty *Phone* i *State* (pierwszy oznacza numer telefonu, raczej nieprzydatny w eksploracji danych, drugi oznacza stan, który, choć przydatny, na potrzeby tego przykładu usuniemy). Uruchom węzeł *Column Filter*.
5. Pozostałe atrybuty nominalne trzeba zamienić na numeryczne. Wykorzystamy w tym celu plik słownikowy *dictchurn.txt*. Zapoznaj się z jego zawartością.
6. Wstaw na przestrzeń roboczą węzeł *File Reader*. Skonfiguruj go do odczytu pliku *dictchurn.txt*. Odznacz opcję *Has RowID*. W zakładce *Transformation* zmień typ dla *Column0* z *Integer* na *String*.
7. Wstaw na przestrzeń roboczą dwa węzły *Manipulation → Column → Convert & Replace → String Replacer (Dictionary)*. Nazwij je odpowiednio *yes_no* i *true_false*.
8. Połącz „górne” wejścia węzłów "szeregowo" do wyjścia węzła *Column Filter* (wyjście *Column Filter* do wejścia *yes_no*, wyjście *yes_no* do wejścia *true_false*). Do „dolnych” wejść węzłów połącz wyjście węzła wczytującego plik *dictchurn.txt.*
9. Skonfiguruj węzły *yes_no* i *true_false*. Parametr *Target Column* ustaw na *Int'l Plan*, *Vmail Plan* dla węzła *yes_no* i *Churn?* dla węzła *true_false.* W obu węzłach wybierz *Column0* dla opcji *Replacement column* oraz *Column1* i *Column2* dla *Pattern column* odpowiednio w węzłach *yes_no* i *true_false.*

10. Na wyjściu węzła *true_false* dodaj nowy węzeł *Manipulation → Column → Convert & Replace → String to Number*. Skonfiguruj węzeł tak, aby uwzględniał atrybuty *Int'l Plan* i *Vmail Plan, Churn?*.
11. Obejrzyj przygotowane dane (możesz w tym celu wykorzystać węzeł *Interactive Table*).

2. Algorytm E-M

1. Na przestrzeń roboczą dołóż węzeł *Analytics → Integrations → Weka → Cluster algorithms → EM*. Na wejście węzła podaj wyjście węzła *String to Number*.
2. Zapoznaj się z opcjami konfiguracyjnymi węzła *EM*. Upewnij się, czy liczba klastrów jest określana automatycznie (wartość -1).
3. Uruchom węzeł *EM* poprzez opcję *Execute and Open Views*.
4. Zapoznaj się z otrzymanymi wynikami.
5. Na przestrzeń roboczą przenieś węzeł *Analytics → Integrations → Weka → Predictors → Weka Cluster Assigner*. Do wejść węzła podłączy odpowiednio wyjścia *String to Number* i *EM*.
6. Uruchom węzeł *Weka Cluster Assigner* za pomocą opcji *Execute*. Z menu kontekstowego wybierz widok *Clustered Data*. Zapoznaj się z wynikami.

3. Algorytm K-średnich

1. Na przestrzeń roboczą przesuń węzeł *Analytics → Mining → Clustering → k-Means*. Zapoznaj się z opisem węzła.
2. Ne wejście węzła *k-Means* podaj wyjście węzła *String to Number*.
3. Zapoznaj się opcjami konfiguracyjnymi węzła *k-Means*. Zwróć uwagę na konieczność podania liczby klastrów. Podaj liczbę klastrów wyznaczoną wcześniej przez algorytm E-M.
4. Uruchom węzeł *k-Means* za pomocą opcji *Execute and Open Views*. Zapoznaj się z wynikami segmentacji i spróbuj znaleźć charakterystyczne cechy klastrów.
5. Zwróć uwagę na rozkłady wartości atrybutów *Int'l Plan*, *Vmail Plan* w wyznaczonych klastrach.
6. Na wyjściu węzła *String to Number* dodaj nowy węzeł *Manipulation → Column → Transform → Normalizer*. Skonfiguruj węzeł tak, aby normalizował wartości wszystkich atrybutów metodą *Min-Max* (0.0-1.0).
7. Na wyjściu węzła *Normalizer* dodaj nowy węzeł *k-Means*. Wejdź do konfiguracji węzła i podaj liczbę klastrów wyznaczoną wcześniej przez algorytm E-M. Uruchom węzeł za pomocą opcji *Execute and Open Views*.
8. Porównaj wyniki segmentacji z wynikami otrzymanymi poprzednio. Jak tym razem rozkładają się (między klastrami) wartości atrybutów *Int'l Plan, Vmail Plan*?