Nowe Trendy w Informatyce

Grupowanie danych niejednorodnych

Skład sekcji: Jacek Wieczorek, Michał Orzeł

08.07.2024r.

Wprowadzenie

W ramach projektu stworzony został program umożliwiający grupowanie danych niejednorodnych. Program może działać w dwóch trybach. W pierwszym z nich, brakujące wartości w danych są uzupełniane bezpośrednio za pomocą algorytmu kNN. W drugim trybie przed uzupełnianiem danych następuje granulacja za pomocą algorytmu FCM. Ma to na celu zredukowanie liczby przykładów dla algorytmu kNN, ponieważ ma on złożoność $O(n^2)$. Celem projektu było zbadanie czy takie podejście istotnie pozwoli zmniejszyć czas wykonania algorytmu przy jednoczesnym zachowaniu jakości grupowania.

Wyniki

Stworzony program został przetestowany na dostarczonych danych wejściowych. Na tej podstawie wykazano, że zastosowanie granulacji pozwala zmniejszyć czas wykonania programu przy zachowaniu podobnej jakości grupowania. Oceny jakości grupowania dokonano ręcznie na podstawie liczby powstałych klastrów oraz wartości wariancji poszczególnych atrybutów w ramach klastra. Czasy wykonania poszczególnych fragmentów algorytmu mierzone zostały za pomocą funkcji dostarczonych w bibliotece `chrono`.

Poniżej przedstawiono wyniki działania programu dla małego zbioru danych. Podejście klasyczne:

```
$ ./ntwi ../dane/male/1/ --seed 1 --granules 3 --clusters 5 --granulation-iters 3
 ↪  --clustering-iters 3 --print-times 1 --print-result 0 --naive 1
t         knn: 0.000324137s
t  clustering: 0.000595384s
t       total: 0.000919521s

ID, Items, Var0, Var1, Var2
0, 30, 22.3867, 3929.92, 0.426096
3, 180, 67.1431, 3327.63, 0.256165
4, 90, 18.4847, 4406.74, 0.638542
```

Algorytm wykorzystujący wstępną granulację (3 granule z każdego źródła danych):

```
$ ./ntwi ../dane/male/1/ --seed 1 --granules 3 --clusters 5 --granulation-iters 3
 ↪  --clustering-iters 3 --print-times 1 --print-result 0 --naive 0
t granulation: 0.000139924s
t         knn: 2.7699e-05s
t  clustering: 0.000106455s
t       total: 0.000274078s

ID, Items, Var0, Var1, Var2
0, 30, 22.3767, 4260.61, 0.550576
```

```
1, 3, 28.7598, 4085.49, 0.488032
4, 57, 54.8432, 3578.23, 0.306535
```

Zbiór danych `male` jest zbyt mały by wykazać różnice w czasie działania obu algorytmów. Można jednak zauważyć, że w obu przypadkach wyniki grupowania są podobne — powstają 3 klastry. Wartości wariancji każdego atrybutu również są zbliżone do siebie.

Działanie algorytmu zostało również sprawdzone na większym dostarczonym zbiorze danych. Z uwagi na obszerność wyników działania programu, zostały one zamieszczone osobno w repozytorium. Na dużym zbiorze danych widać jednak bardzo wyraźnie różnice w czasie działania algorytmu — ponad 300-krotne przyspieszenie, przy zachowaniu podobnej liczby klastrów i podobnych wartości wariancji.

Podejście klasyczne:

```
t         knn: 141.844s
t  clustering: 8.87688s
t       total: 150.721s
```

Podejście z granulacją:

```
t granulation: 0.434979s
t         knn: 0.000260982s
t  clustering: 0.0124396s
t       total: 0.44768s
```

Wnioski

W ramach projektu wykazano, że zaproponowany algorytm grupowania z wstępną granulacją może być bardzo opłacalny czasowo. Wyniki algorytmu zależą też od wielu parametrów — począwszy od wartości k dla algorytmu kNN, przez liczbę iteracji granulacji i grupowania aż do liczby pożądanych klastrów i granuli. Ciekawym usprawnieniem mogłoby być opracowanie metody automatycznego doboru tych parametrów. Warto zaznaczyć, że aktualna implementacja algorytmu pozostawia także duże pole dla usprawnień i optymalizacji wydajności.