Streszczenie

Tytuł rozprawy: *Adaptacyjne algorytmy hierarchicznej analizy skupień oparte na metodach agregacji danych*
Autor: Anna Cena
Promotor: dr hab. inż. Marek Gągolewski, prof. PAN

Analiza skupień (ang. *cluster analysis*), ma na celu wyznaczenie podziału wejściowego zbioru danych tak, by obserwacje w obrębie jednej grupy były jak najbardziej podobne do siebie względem wybranego kryterium, zaś obserwacje należące do osobnych grup różniły się między sobą możliwie istotnie. Innymi słowy, celem analizy skupień jest identyfikacja naturalnej, ukrytej *struktury* danych. Metody analizy skupień są szeroko wykorzystywane m.in. w naukach biologicznych, społecznych, przetwarzaniu obrazów i innych sygnałów itd. Co istotne, w przypadku tego zadania najczęściej nie posiadamy żadnej wiedzy dotyczącej poprawnego grupowania – algorytmy analizy skupień opierają się więc jedynie na ocenie „podobieństwa" lub „bliskości" poszczególnych elementów zbioru (mierzonej przy użyciu szeroko rozumianej funkcji odległości) lub relacji sąsiedztwa, czy też badaniu zagęszczenia obserwacji.

Wśród algorytmów analizy skupień warto wyróżnić *metody hierarchiczne*, które w wyniku zwracają całą hierarchię – rodzinę zagnieżdżonych podziałów – taką, że każdy jej element jest sam w sobie poprawnym grupowaniem. Warto w tym miejscu zwrócić uwagę, że wynikowa hierarchia cechuje się spójnością grupowań na każdym ze swych poziomów – podział na danym poziomie hierarchii składa się zawsze z grup wzajemnie zawierających się lub tożsamych ze skupieniami, które formują podział występujący „wyżej" w hierarchii. Metody tego typu oferują więc nie tylko wgląd w strukturę zbioru danych, ale także możliwość obserwowania procesu formowania skupień.

Co więcej, większość metod hierarchicznych nie wymaga wielu założeń dotyczących analizowanego zbioru: najczęściej wymagamy jedynie, by określona była funkcja odległości między punktami, która może już uwzględniać efekt wstępnego przetwarzania danych – np. ważenie czy ekstrakcję cech, oraz jej rozszerzenie na odległość między skupieniami (tzw. odmienność). W zależności od doboru odmienności, algorytmy te mogą brać pod uwagę lokalną „gęstość" rozkładu danych jak i podobieństwa na ogólniejszym poziomie. W niniejszej rozprawie przeprowadzono szczegółową analizę odmienności opartej na operatorach OWA. Przypomnijmy, że funkcje OWA uogólniają niektóre z najpopularniejszych odmienności wykorzystywanych w algorytmach aglomeracyjnych tj. odmienność najbliższego i najdalszego sąsiada oraz odmienność średnią (określone, odpowiednio, jako minimum, maksimum oraz średnią arytmetyczną odległości między skupieniami). Stąd ich wykorzystanie w hierarchicznej analizie skupień, zaproponowane oryginalnie przez Yagera (2000), a następnie Nasıbova i Kandemır-Cavas (2011), jest naturalnym rozszerzeniem klasycznego podejścia. Po pierwsze, zaproponowano tutaj szeroką gamę zestawów wag, które następnie wykorzystano do analizy zbiorów benchmarkowych. Uwzględniono zarówno kwantyle próbkowe odległości i funkcje OWA je „rozmywające", jak i różne inne typy średnich. Dokonano również rozszerzenia odmienności OWA, polegającego na rozbiciu procesu agregacji na trzy funkcje składowe. Następnie zbadano wpływ doboru

wag funkcji OWA na wyniki grupowania.

Niestety algorytmy hierarchiczne są dość kosztowne obliczeniowo – wyjątek stanowi tu algorytm z odmiennością najbliższego sąsiada, który może zostać zaimplementowany w oparciu o minimalne drzewo rozpinające. Wadą tego podejścia jest niestety podatność na obserwacje odstające oraz skłonność do generowania podziałów bardzo niezrównoważonych. Algorytm Genie, oparty na kryterium *single linkage*, zachowuje możliwość szybkiego działania, redukując jednocześnie wady otrzymanej w wyniku hierarchii. Metoda Genie zawiera korektę na nierówność rozkładu liczności skupień, tj. algorytm wymusza wcześniejsze łączenie skupień o minimalnej liczności (potencjalnie obserwacji odstających), jeśli stopień nierówności rozkładu przekroczy pewien ustalony próg g. W niniejszej rozprawie przeprowadzono szeroką analizę jakości podziałów uzyskanych przy użyciu algorytmu Genie na zbiorach benchmarkowych. Zweryfikowano również wpływ zastosowania zaproponowanej w nim poprawki na działanie algorytmu aglomeracyjnego z odmiennościami OWA. Wykazano, że algorytm Genie nie tylko znacząco poprawia jakość generowanych podziałów (mierzoną jako zgodność z etykietami referencyjnymi – w porównaniu do klasycznych algorytmów aglomeracyjnych oraz analizowanych odmienności OWA), ale także odnotowano, że zastosowanie poprawki Genie w algorytmach z odmiennościami OWA powoduje znaczący wzrost jakości generowanych podziałów. Dokonano również oceny wrażliwości parametrów algorytmu Genie: miary określającej stopień nierówności rozkładu liczności skupień oraz progu odcięcia g. Sformułowano również rekomendację dotyczącą ich wartości.

W niniejszej rozprawie zaproponowano także autorską metodę aglomeracyjną hierarchicznej analizy skupień opartą na minimalnym drzewie rozpinającym. Przedstawione tu podejście wykorzystuje częściowe informacje na temat struktury danych uzyskane przy użyciu algorytmu Genie – uwzględnione w postaci podziału inicjującego hierarchię. Grupowanie startowe zostaje utworzone adaptacyjnie jako przecięcie podziałów wygenerowanych przy użyciu algorytmu Genie dla różnych wartości progów odcięcia korekty na rozkład liczności skupień. Dodatkowo rozpatrzono sposób wyboru podziału oparty na kryterium informacyjnym, a także w postaci problemu minimalizacji odmienności najbliższego sąsiada z pewnymi ograniczeniami. Następnie przeprowadzono porównanie zaproponowanej metody: oryginalnego algorytmu Genie oraz dostępnych w literaturze metod analizy skupień, w tym klasycznych algorytmów aglomeracyjnych, algorytmu przez dzielenie wykorzystującego kryterium informacyjne, metody K-średnich, HDBSCAN*, Birch, metody spektralnej itd. Uzyskane wyniki wskazały na wysoką efektywność zaproponowanego algorytmu – wygenerowane przy jego użyciu podziały wykazują się największą, spośród analizowanych metod, zgodność z podziałami referencyjnymi. Co więcej, zaproponowana metoda nie wymaga specyfikacji dodatkowych parametrów.