

Jörn Lötsch, Alfred Ultsch "Comparative assessment of projection and clustering method combinations in the analysis of biomedical data"

Celem artykułu jest rzetelna ocena skuteczności popularnych połączeń metod redukcji wymiaru i klastrowania, stosowanych w analizie danych biomedycznych. Autorzy zwracają uwagę na powszechną praktykę polegającą na tym, że dane wielowymiarowe najpierw rzutuje się do przestrzeni o mniejszym wymiarze (np. dla wizualizacji lub uproszczenia struktury), a następnie wykonuje się klastrowanie na tak uzyskanej reprezentacji. Jednocześnie wskazują, że cele projekcji i klastrowania mogą być rozbieżne: klasyczne PCA koncentruje się na uchwyceniu kierunków największej wariancji, natomiast klastrowanie poszukuje lokalnych zagęszczeń i sąsiedztw. W efekcie popularny schemat PCA jako krok wstępny nie musi być optymalny i wymaga empirycznej weryfikacji.

Badane zagadnienie ma charakter porównawczy: autorzy chcą sprawdzić, czy istnieją kombinacje metod, które w sposób stabilny i powtarzalny odtwarzają sensowną strukturę grup w różnych typach danych, oraz czy można wskazać bezpieczną, domyślną strategię postępowania.

W badaniu wykorzystano łącznie czternaście zbiorów danych: dziewięć zbiorów sztucznych oraz pięć zbiorów rzeczywistych z obszaru biomedycyny. Taki dobór ma uzasadnienie metodyczne: dane sztuczne pozwalają kontrolować strukturę (i łatwiej interpretować, co metoda powinna wykryć), natomiast dane rzeczywiste odzwierciedlają typowe trudności spotykane w praktyce, takie jak szum, nieliniowość i niejednoznaczne granice między grupami.

Artykuł jest zbudowany jako przegląd i test wielu wariantów, ale w kontekście bloku II kluczowe są cztery elementy: dwie metody projekcji (PCA, MDS) oraz dwie metody klastrowania (k-means, k-medoids realizowane algorytmem z rodziny PAM).

PCA pełni rolę metody bazowej do redukcji wymiaru. Jest to najczęściej spotykany wybór w analizie danych i dlatego autorzy traktują go jako naturalny punkt odniesienia w eksperymentach.

MDS jest ujęte jako alternatywny sposób projekcji, w którym istotą jest możliwie dobre zachowanie relacji podobieństwa lub odległości między obserwacjami po rzutowaniu do mniejszego wymiaru. W praktyce MDS bywa stosowane, gdy zależy na interpretacji w kategoriach odległości pomiędzy obiektami, a nie tylko na maksymalizacji wariancji.

W części klastrowania autorzy uwzględniają k-means jako standardową metodę centroidową. Zastosowanie k-means jest uzasadnione tym, że jest to algorytm powszechny, szybki i często używany jako pierwszy wybór w zadaniach segmentacji danych.

Drugą metodą klastrowania istotną dla tematu bloku II jest k-medoids, które w praktyce jest najczęściej realizowane algorytmem PAM. Różnica względem k-means polega na tym, że reprezentant klastra (medoid) jest rzeczywistą obserwacją z danych, co często zwiększa odporność na obserwacje odstające oraz pozwala pracować na ogólnych macierzach odległości, a nie wyłącznie na średnich w przestrzeni euklidesowej.

W eksperymencie testowane są kombinacje typu: projekcja (PCA lub MDS) a później klastrowanie (k-means lub k-medoids/PAM). Autorzy porównują wyniki między zbiorami danych i między parametryzacjami, a następnie oceniają jakość zarówno ilościowo, jak i jakościowo na wykresach.

Najważniejszym wynikiem pracy jest brak uniwersalnej kombinacji projekcji i klastrowania, która konsekwentnie odtwarzałaby sensowną strukturę grup we wszystkich testowanych zbiorach. Autorzy

wprost wskazują, że żadna z badanych konfiguracji nie osiąga stabilnie dobrych rezultatów w każdym przypadku. Taki wniosek ma dużą wartość praktyczną, ponieważ podważa przekonanie, że istnieje standardowy, bezpieczny schemat postępowania niezależny od danych.

Drugi istotny wniosek dotyczy PCA. Autorzy dochodzą do konkluzji, że wyniki nie uzasadniają traktowania PCA jako domyślnej projekcji przed klastrowaniem. W części porównawczej wskazują, że PCA bywało dorównywane lub przewyższane przez metody oparte na sąsiedztwach i uczeniu rozmaitości (np. UMAP, t-SNE, isomap), choć nie w sposób całkowicie jednoznaczny. Z perspektywy Przekłada się to na prostą zasadę: wybór redukcji wymiaru powinien wynikać z charakteru danych i celu analizy a nie z przyzwyczajenia.

Dodatkową wartością artykułu jest podkreślenie znaczenia weryfikacji wizualnej. Autorzy argumentują, że sama metryka liczbowa nie zawsze oddaje, czy uzyskany podział jest interpretowalny i zgodny z oczekiwaną strukturą, dlatego proponują sposób wizualizacji oparty na diagramach Woronoja i kodowaniu kolorami, który ma ułatwiać ocenę granic klastrów w przestrzeni po projekcji. W praktyce analitycznej jest to ważne, ponieważ dobór metody często odbywa się iteracyjnie i łączy kryteria ilościowe z oceną interpretowalności wyniku.

Szeroki zakres porównania jest jednocześnie zaletą i ograniczeniem. Artykuł pokazuje, że nie ma jednego rozwiązania dla wszystkich danych, ale przez to nie dostarcza jednoznacznej, prostej instrukcji wyboru metod dla konkretnego przypadku. Dla odbiorcy oczekującego rekomendacji typu najlepsza metoda w większości zadań, wnioski mogą być mniej satysfakcjonujące, ponieważ podkreślają zależność od danych.