StreamlitDRV: Narzędzie do wizualizacji danych z wykorzystaniem Streamlit

Sprawozdanie z projektu

Wojciech Bartoszek
Bartosz Gacek
Jarosław Kołdun

2025

1 Wprowadzenie

StreamlitDRV to interaktywna aplikacja webowa służąca do wizualizacji danych poprzez zastosowanie metod redukcji wymiarowości. Projekt został zrealizowany z wykorzystaniem frameworka Streamlit, który umożliwia szybkie tworzenie aplikacji analitycznych z interfejsem webowym. Głównym celem aplikacji jest umożliwienie użytkownikom eksploracji własnych zbiorów danych oraz porównania różnych technik redukcji wymiarowości w intuicyjny sposób.

Aplikacja adresuje potrzebę wizualizacji wielowymiarowych zbiorów danych, które w swojej pierwotnej formie są trudne do interpretacji. Poprzez redukcję do dwóch wymiarów, użytkownicy mogą odkrywać ukryte wzorce i struktury w danych, co ma szczególne znaczenie w analizie eksploracyjnej i prezentacji wyników badań.

2 Architektura systemu

Projekt został zaprojektowany zgodnie z zasadami modularności i separacji odpowiedzialności, co zapewnia czytelność kodu, łatwość rozszerzania funkcjonalności oraz efektywne zarządzanie złożonością systemu.

2.1 Struktura modułowa

Architektura składa się z głównego pliku aplikacji `app.py` oraz sześciu wyspecjalizowanych modułów znajdujących się w katalogu `src/`. Każdy moduł ma jasno zdefiniowaną odpowiedzialność i interfejs, co umożliwia niezależny rozwój i testowanie poszczególnych komponentów.

Moduł `config.py` zawiera konfigurację aplikacji, opisy metod oraz wytyczne dotyczące doboru parametrów, centralizując wszystkie ustawienia systemowe. `data_handler.py` od-

powiada za ładowanie, walidację i preprocessing danych, obsługując zarówno wbudowany zbiór danych, jak i pliki przesłane przez użytkownika. `reduction_methods.py` implementuje wszystkie dostępne metody redukcji wymiarowości wraz z ich konfigurowalnymi parametrami.

Wizualizacja danych jest realizowana przez moduł `visualizations.py`, który wykorzystuje bibliotekę Plotly do tworzenia interaktywnych wykresów. Moduł `metrics.py` dostarcza narzędzi do oceny jakości redukcji wymiarowości, obliczając różnorodne metryki i przeprowadzając analizę wariancji. Ostatni moduł, `parameter_optimization.py`, implementuje funkcjonalność optymalizacji parametrów poprzez przeszukiwanie siatki oraz wizualizację wyników w postaci map ciepła.

2.2 Zarządzanie zależnościami i środowisko

Do zarządzania zależnościami wykorzystano nowoczesny menedżer pakietów `uv`, który stanowi szybką alternatywę dla pip i poetry. `uv` zapewnia deterministyczne instalowanie pakietów, znacznie szybsze operacje oraz lepsze zarządzanie środowiskami wirtualnymi. Konfiguracja projektu jest zdefiniowana w pliku `pyproject.toml`, co umożliwia standardowe zarządzanie metadanymi i zależnościami zgodnie z PEP 621.

Projekt wykorzystuje blokowanie wersji zależności poprzez plik `uv.lock`, co gwarantuje reprodukowalność środowiska na różnych maszynach i w różnych fazach rozwoju projektu. To podejście eliminuje problemy związane z niekompatybilnością wersji bibliotek i zapewnia stabilność działania aplikacji.

2.3 Konteneryzacja i wdrażanie

Projekt został dodatkowo przygotowany do konteneryzacji przy użyciu Docker, co zapewnia spójność środowiska uruchomieniowego niezależnie od platformy docelowej. Dockerfile definiuje kompletne środowisko aplikacji, włączając w to wszystkie zależności systemowe i konfigurację Pythona.

Dzięki konteneryzacji aplikacja może być łatwo wdrażana w różnych środowiskach - od lokalnych maszyn deweloperskich po serwery produkcyjne - bez konieczności konfigurowania zależności lokalnie. Kontener zapewnia również izolację aplikacji od systemu hosta, co zwiększa bezpieczeństwo i stabilność wdrożenia.

3 Implementowane metody redukcji wymiarowości

Aplikacja oferuje sześć różnych metod redukcji wymiarowości, każda z własnymi charakterystykami, parametrami konfiguracyjnymi i zastosowaniami. Wybór odpowiedniej metody zależy od charakteru danych, wymagań dotyczących zachowania struktury oraz ograniczeń obliczeniowych.

3.1 Principal Component Analysis (PCA)

Analiza Głównych Składowych stanowi klasyczną metodę liniową redukcji wymiarowości, która znajduje kierunki maksymalnej wariancji w danych poprzez dekompozycję macierzy kowariancji. PCA jest deterministyczna i efektywna obliczeniowo, co czyni ją idealną do

wstępnej analizy danych oraz przypadków wymagających szybkiego przetwarzania dużych zbiorów danych.

Metoda zakłada liniowe relacje między zmiennymi i jest szczególnie skuteczna dla danych o wysokiej korelacji między cechami. W aplikacji PCA służy również jako punkt odniesienia dla porównania z metodami nieliniowymi, umożliwiając ocenę, czy nieliniowość danych uzasadnia zastosowanie bardziej złożonych algorytmów.

3.2 Kernel Principal Component Analysis (KPCA)

Kernel PCA rozszerza możliwości klasycznej PCA o nieliniowe transformacje poprzez zastosowanie funkcji jądra, które mapują dane do przestrzeni o wyższej wymiarowości, gdzie relacje liniowe mogą lepiej reprezentować nieliniowe struktury w oryginalnej przestrzeni.

Aplikacja implementuje cztery typy jąder: RBF (Radial Basis Function) skuteczne dla danych o lokalnych skupiskach, jądra wielomianowe odpowiednie dla danych o strukturze algebraicznej, jądra sigmoidalne naśladujące funkcje aktywacji sieci neuronowych oraz jądra cosinusowe wykorzystujące miary podobieństwa kątowego. Parametr gamma kontroluje wpływ pojedynczych punktów treningowych, gdzie wyższe wartości prowadzą do bardziej złożonych granic decyzyjnych.

3.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE to probabilistyczna metoda nieliniowa szczególnie skuteczna w wizualizacji skupisk danych poprzez zachowanie lokalnej struktury sąsiedztwa. Algorytm modeluje podobieństwa między punktami jako prawdopodobieństwa i minimalizuje rozbieżność Kullbacka-Leiblera między rozkładami w przestrzeni oryginalnej i zredukowanej.

Kluczowym parametrem jest perplexity, który kontroluje liczbę najbliższych sąsiadów uwzględnianych w analizie. Niskie wartości perplexity (5-15) podkreślają bardzo lokalne struktury, podczas gdy wyższe wartości (30-50) zachowują bardziej globalne relacje. Metoda jest szczególnie skuteczna w odkrywaniu ukrytych skupisk, ale może być wrażliwa na wybór parametrów i kosztowna obliczeniowo dla dużych zbiorów danych.

3.4 Uniform Manifold Approximation and Projection (UMAP)

UMAP opiera się na teorii rozmaitości topologicznych i oferuje kompromis między zachowaniem struktury lokalnej i globalnej przy jednoczesnej wysokiej wydajności obliczeniowej. Algorytm konstruuje graf sąsiedztwa w przestrzeni oryginalnej, a następnie optymalizuje podobną strukturę w przestrzeni zredukowanej.

Parametr `n_neighbors` kontroluje równowagę między strukturą lokalną a globalną - więcej sąsiadów prowadzi do lepszego zachowania struktury globalnej kosztem szczegółów lokalnych. Parametr `min_dist` określa minimalne odległości między punktami w przestrzeni zredukowanej, gdzie niższe wartości tworzą bardziej zwarte skupiska. UMAP jest szczególnie skuteczny dla danych o złożonej strukturze topologicznej.

3.5 TriMap

TRIMAP wykorzystuje tripletowe relacje między punktami do zachowania zarówno lokalnych sąsiedztw, jak i globalnych odległości. Algorytm definiuje zbiory tripletów składają-

cych się z punktu centralnego oraz jego najbliższych i najdalszych sąsiadów, a następnie optymalizuje reprezentację w przestrzeni zredukowanej tak, aby zachować te relacje.

Parametr `n_inliers` określa liczbę najbliższych sąsiadów uwzględnianych dla każdego punktu, wpływając na zachowanie lokalnej struktury. Parametr `n_outliers` kontroluje liczbę odległych punktów używanych do zachowania struktury globalnej. TRIMAP często zapewnia lepsze zachowanie struktury globalnej niż t-SNE przy porównywalnej jakości reprezentacji lokalnej.

3.6 Pairwise Controlled Manifold Approximation Projection (PaCMAP)

PaCMAP kontroluje zachowanie struktury lokalnej i globalnej poprzez precyzyjne zarządzanie trzema typami par punktów: bliskimi parami (Near Pairs) zachowującymi strukturę lokalną, średnimi parami (Mid-Near Pairs) łączącymi regiony lokalne oraz dalekimi parami (Far Pairs) zachowującymi strukturę globalną.

Parametr `MN_ratio` kontroluje proporcję par średnich do par bliskich, wpływając na połączenia między lokalnymi regionami. Parametr `FP_ratio` określa stosunek par dalekich do par bliskich, kontrolując zachowanie struktury globalnej. Ta precyzyjna kontrola czyni PaCMAP szczególnie skutecznym w zachowaniu zarówno lokalnych skupisk, jak i globalnej organizacji danych.

4 Funkcjonalności aplikacji

Aplikacja oferuje kompleksowy i intuicyjny przepływ pracy, który prowadzi użytkownika przez wszystkie etapy analizy danych - od wczytania danych, przez konfigurację parametrów, aż po szczegółową analizę wyników. Interfejs został zaprojektowany tak, aby być dostępny zarówno dla początkujących użytkowników, jak i ekspertów wymagających zaawansowanych opcji konfiguracji.

4.1 Zarządzanie danymi i preprocessing

System obsługuje dwa główne źródła danych: wbudowany zbiór danych dotyczący cukrzycy, który służy jako przykład demonstracyjny, oraz możliwość przesłania własnych plików w formatach CSV i Excel. Dla przesyłanych plików aplikacja automatycznie wykrywa typy kolumn i umożliwia użytkownikowi wybór zmiennej docelowej oraz cech do analizy.

Moduł preprocessingu implementuje kompleksowy pipeline przygotowania danych. System automatycznie identyfikuje i obsługuje brakujące wartości poprzez usuwanie wierszy z niepełnymi danymi, co zapewnia stabilność algorytmów redukcji wymiarowości. Wszystkie dane numeryczne są standaryzowane przy użyciu StandardScaler, co jest kluczowe dla poprawnego działania metod opartych na odległościach.

Dla dużych zbiorów danych aplikacja oferuje funkcjonalność próbkowania, która umożliwia ograniczenie liczby obserwacji w celu przyspieszenia obliczeń. Użytkownik może określić maksymalną liczbę próbek, a system zastosuje losowe próbkowanie stratyfikowane zachowujące proporcje klas w zmiennej docelowej.

4.2 Interfejs konfiguracji i kontrola parametrów

Panel boczny aplikacji zawiera dynamicznie generowane kontrolki parametrów dostosowane do wybranej metody redukcji wymiarowości. Każda metoda ma zdefiniowany zestaw kluczowych parametrów z odpowiednimi zakresami wartości i wartościami domyślnymi wybranymi na podstawie najlepszych praktyk.

System implementuje inteligentne ostrzeżenia wydajnościowe, które informują użytkownika o potencjalnie długich czasach obliczeń dla kosztownych metod (t-SNE, TRIMAP) zastosowanych do dużych zbiorów danych. Ostrzeżenia zawierają konkretne sugestie alternatywnych podejść, takich jak próbkowanie danych lub wybór szybszych metod.

Aplikacja dostarcza również szczegółowe wytyczne dotyczące doboru parametrów dla każdej metody, wyjaśniając wpływ poszczególnych parametrów na wyniki i podając rekomendacje dla różnych typów danych i celów analizy.

4.3 Wizualizacja i prezentacja wyników

Głównym elementem prezentacji wyników są interaktywne wykresy rozproszenia tworzone przy użyciu biblioteki Plotly. Punkty są kolorowane zgodnie ze zmienną docelową, co umożliwia natychmiastową ocenę jakości separacji klas w przestrzeni zredukowanej. Wykresy oferują pełną interaktywność, w tym powiększanie, przesuwanie oraz wyświetlanie szczegółowych informacji o punktach.

System automatycznie dostosowuje skalę kolorów do typu zmiennej docelowej - dla zmiennych kategorycznych używa dyskretnej palety kolorów, podczas gdy dla zmiennych ciągłych stosuje gradient kolorystyczny. Wykresy zawierają również odpowiednie legendy i opisy osi.

4.4 Analiza jakości i metryki

Aplikacja implementuje rozbudowany system oceny jakości redukcji wymiarowości. Dla metod liniowych, takich jak PCA, obliczana jest analiza wariancji wyjaśnionej, która pokazuje, jaka część całkowitej wariancji danych została zachowana w reprezentacji dwuwymiarowej.

System oblicza również metryki rekonstrukcji, które mierzą, jak dobrze oryginalne dane mogą być odtworzone z reprezentacji zredukowanej. Obejmuje to błąd średniokwadratowy rekonstrukcji oraz wizualizacje rozkładu błędów, które pomagają zidentyfikować punkty szczególnie trudne do reprezentacji w niższej wymiarowości.

Dla wszystkich metod dostępna jest analiza zachowania odległości, która porównuje odległości między punktami w przestrzeni oryginalnej i zredukowanej, pomagając ocenić, jak dobrze metoda zachowuje strukturę danych.

4.5 Optymalizacja parametrów

Zaawansowana funkcjonalność optymalizacji parametrów umożliwia automatyczne przeszukiwanie przestrzeni parametrów w celu znalezienia optymalnych ustawień dla danego zbioru danych. System implementuje przeszukiwanie siatki (grid search) dla kluczowych parametrów każdej metody.

Wyniki optymalizacji są prezentowane w postaci interaktywnych map ciepła, które wizualizują jakość wyników dla różnych kombinacji parametrów. Mapy ciepła używają różnych metryk jakości, takich jak zachowanie odległości lokalnych czy globalna struktura danych, umożliwiając użytkownikowi wybór parametrów optymalnych dla konkretnego celu analizy.

System automatycznie identyfikuje i wyróżnia kombinacje parametrów dające najlepsze wyniki, a także dostarcza szczegółowe wyjaśnienia dotyczące interpretacji wyników optymalizacji.

5 Aspekty techniczne

Projekt wykorzystuje nowoczesne narzędzia do zarządzania zależnościami, w szczególności menedżer pakietów `uv`, który zapewnia szybką instalację i deterministyczne środowisko. Aplikacja została skonteneryzowana przy użyciu Docker, co ułatwia wdrażanie i zapewnia spójność środowiska uruchomieniowego.

Kluczowe biblioteki obejmują Streamlit jako framework webowy, scikit-learn dla klasycznych metod uczenia maszynowego, oraz specjalistyczne implementacje UMAP, TRIMAP i PaCMAP. Wizualizacje są realizowane przy użyciu Plotly, co zapewnia interaktywność i wysoką jakość graficzną.

Aplikacja implementuje mechanizmy optymalizacji wydajności, w tym ostrzeżenia dla dużych zbiorów danych, możliwość próbkowania oraz inteligentne dobieranie domyślnych parametrów. Obsługa błędów i walidacja danych zapewniają stabilność działania w różnych scenariuszach użycia.

6 Wnioski i możliwości rozwoju

StreamlitDRV stanowi kompleksowe narzędzie do eksploracji danych poprzez redukcję wymiarowości, łączące łatwość użycia z zaawansowanymi możliwościami analitycznymi. Modularność architektury umożliwia łatwe rozszerzanie funkcjonalności o nowe metody lub metryki oceny.

Projekt może być rozwijany w kierunku implementacji dodatkowych metod redukcji wymiarowości, rozszerzenia możliwości analizy statystycznej oraz integracji z zewnętrznymi źródłami danych. Potencjalne ulepszenia obejmują również implementację redukcji do więcej niż dwóch wymiarów oraz możliwość eksportu wyników w różnych formatach.

Aplikacja może znaleźć zastosowanie w edukacji, badaniach naukowych oraz analizie danych biznesowych, oferując dostępne narzędzie do wizualizacji i eksploracji wielowymiarowych zbiorów danych.