Przetwarzanie danych wysokowymiarowych za pomocą kompresji próbek i redukcji wymiarowości

Wiele nowoczesnych zastosowań informatyki wiąże się z przechowywaniem i przetwarzaniem ogromnych ilości danych wysokowymiarowych. Projektowanie algorytmów i struktur danych w celu wydajnej obsługi dużych ilości danych wysokowymiarowych jest zatem kluczowym wyzwaniem stojącym dziś przed projektantami algorytmów i programistami.

Rozważmy na przykład aplikację do wyszukiwania obrazów, która pobiera obraz jako dane wejściowe i zwraca „najbliższe dopasowanie" z bazy danych obrazów. Zakładając, że każdy obraz ma rozdzielczość 48 megapikseli i Truecolor, oznacza to, że pojedynczy obraz składa się z 48 milionów pikseli, przy czym każdy piksel ma jedną z 16 milionów możliwych wartości kolorów. Zatem każdy obraz jest punktem w przestrzeni o 48 milionach wymiarów, a wyszukiwanie obrazów oznacza po prostu wyszukiwanie najbliższego sąsiada punktu wejściowego (tj. obrazu). Choć w praktyce obraz jest zazwyczaj zapisywany po wstępnej obróbce w celu wydobycia mniejszej liczby „cech", to ich liczba wciąż może sięgać setek, a nawet tysięcy, co oznacza, że algorytm nadal musi pracować w kilkuset wymiarach.

Ważną techniką rozwiązywania tych problemów jest koncepcja wstępnego przetwarzania, co oznacza, że w przypadku zbioru danych o dużej objętości i wysokowymiarowości chcemy zmniejszyć jego rozmiar i/lub wymiarowość przed przedstawieniem go algorytmowi przetwarzania danych. Zatem możemy ogólnie podzielić takie techniki przetwarzania wstępnego na dwie klasy – te, które zmniejszają rozmiar zbioru danych lub techniki kompresji próbki, oraz te, które zmniejszają wymiarowość, czyli techniki redukcji wymiarowości. Naturalnie w obu przypadkach chcielibyśmy, aby taka redukcja możliwie najlepiej zachowała zawartość informacyjną danych pierwotnych.

Takie techniki wstępnego przetwarzania są stosowane od lat 80. XX wieku i nadal ewoluują wraz z postępem technologii komputerowej. Jednak większość tych metod została pierwotnie zaprojektowana w celu zachowania informacji związanych z odległościami euklidesowymi lub bliskością między punktami. W ciągu ostatnich dwudziestu lat sukces sieci neuronowych, uczenia maszynowego i sztucznej inteligencji wprowadził nowe narzędzia analizy danych, w tym statystyczne funkcje odległości i miary odmienności. Dodatkowo, właściwości geometryczne danych wyższego rzędu stały się bardziej znaczące. Wiele konwencjonalnych metod kompresji próbek i redukcji wymiarów może nie być kompatybilnych z tymi miarami lub ich nie uwzględniać.

Naszym celem jest ocena istniejących technik obróbki wstępnej pod kątem miar odmienności i właściwości geometrycznych, przy jednoczesnym opracowaniu nowych metod, jeśli zajdzie taka potrzeba. Te nowe techniki zapewnią matematycznie udowodnione gwarancje ich działania. Na przykład w przypadku losowych rzutów redukcji wymiarowości zbadamy ich zachowanie w odniesieniu do cech geometrycznych wyższego rzędu i stworzymy metody redukcji wymiarowości kompatybilne z miarami odmienności, a nie tylko odległościami euklidesowymi. Podobnie w przypadku kompresji próbek naszym celem jest przystosowanie teorii VC do pracy z nowymi miarami odmienności oraz zbadanie technik topologicznych i geometrii dyskretnej w celu rozwiązania trudnych przypadków. Na koniec zaprojektujemy ulepszone algorytmy minimalizacji rozbieżności w celu uzyskania rzadkich próbek przy użyciu iterowanych zrównoważonych partycji.

Główne wyniki projektu obejmują (i) nowatorskie techniki redukcji wymiarowości, (ii) matematyczne zapewnienia dotyczące skuteczności ustalonych metod obsługi nowych typów danych, (iii) ulepszone limity kompresji próbek dla systemów zgodnych z warunkami ograniczonych wymiarów VC oraz (iv) innowacyjne metody kompresji próbek dla systemów o nieograniczonym wymiarze VC, wykorzystujące dyskretną geometrię, topologię i kombinatorykę. Dodatkowo naszym celem jest osiągnięcie (v) ulepszonych algorytmów minimalizacji rozbieżności. Celem projektu jest także wspieranie powiązań między algorytmami wysokowymiarowymi a różnymi gałęziami teorii, takimi jak algorytmy losowe, prawdopodobieństwo wysokowymiarowe, wykrywanie skompresowane i analiza geometryczna.