Nr rejestracyjny: 2018/29/N/ST6/00676; Kierownik projektu: mgr Piotr Wygocki

Problem najbliższych sąsiadów bez fałszywie ujemnych wyników – streszczenie popularnonaukowe.

Piotr Wygocki

5 czerwca 2018

1 Motywacja

Załóżmy, że mamy daną bardzo dużą kolekcję obrazków i każdy z nich jest reprezentowany przez krotkę setek liczb rzeczywistych. Takie dane wejściowe mogą być przetworzone w celu utworzenia struktury danych, która posłuży do odpowiedzi na pytania. Pytanie to kolejny obrazek, a zadaniem jest znalezienie obrazka ze zbioru wejściowego, który jest najbardziej podobny do obrazka z zapytania. Taka technika może być użyta np. do klasyfikacji obrazków. Załóżmy, że chcemy przypisać dany obrazek do jednej z danych kategorii. Nasza strategia może polegać na znalezieniu najbliższego sąsiada z już skategoryzowanego zestawu danych wejściowych i przypisania do zapytania takiej samej kategorii jaką ma znaleziony najbliższy sąsiad.

Problem opisany powyżej nazywa się problemem najbliższego sąsiada. Jedna z metod na znalezienie najbliższego sąsiada to przejrzenie całego zbioru wejściowego, obliczenie miary podobieństwa i zwrócenie najbliższego punktu. Niestety ta metoda jest niedopuszczalna w wielu praktycznych zastosowaniach, gdzie odpowiedź na pytanie musi być wydajna. Problem można rozwiązać, gdy wymiar przestrzeni jest niski. Niestety, przy silnych założeniach teoretycznych, problem w przestrzeniach o wysokim wymiarze nie może być efektywnie rozwiązany, tj. z czasem zapytania podliniowym w liczbie punktów wejściowych i zarówno czasem zapytania, jak i czasem pre–processingu wielomianowym ze względu na wymiar przestrzeni. Rozważa się więc przybliżony problem najbliższego sąsiada (ang. ANN). Większość wcześniej znanych algorytmów dla ANN dawała jedynie gwarancje Monte Carlo, tj. algorytm może zwrócić niepoprawną odpowiedź z pewnym prawdopodobieństwem.

2 Cel

Naszym głównym celem jest zaprojektowanie wydajnego algorytmu z mocniejszymi gwarancjami, np. z gwarancjami Las Vegas. Oznacza to, że wynikiem algorytmu jest zawsze najbliższy sąsiad. Istniejące algorytmy często mają wiele trudnych do dostrojenia parametrów. W tym projekcie planujemy rozwiązać ten problem, co może wpłynąć na różne obszary nauki, w których używa się ANN, na przykład w rozpoznawaniu obrazów, robotyce lub rozpoznawaniem wzorców.

3 Badania

Badania skupią się na zastosowaniu różnych technik z obszaru geometrii obliczeniowej. Problem najbliższego sąsiada jest dobrze zrozumiany w przypadku przestrzeni niskowymiarowych. Ponadto istnieją wydajne algorytmy, nawet jeśli wymiar przestrzeni wejściowej jest rzędu logarytmu liczby punktów wejściowych. Naturalnym pomysłem jest zmniejszyć wymiar przestrzeni wejściowej i rozwiązać problem niskowymiarowy. Takie redukcje są często implementowanie jako rzutowanie przestrzeni liniowych. Niestety, takie rzutowania są zwykle probabilistyczne – elementy macierzy rzutowania są losowane niezależnie z pewnego rozkładu prawdopodobieństwa. Co więcej, te przekształcenia, zachowują odległości tylko z pewnym prawdopodobieństwem (patrz np. Lemat Johnsona Lindenstraussa). Istnieją wyniki derandomizacyjne, ale z dokładnością do wiedzy autora, żadne z nich nie mogą być bezpośrednio zastosowane w ANN. Badania będą skoncentrowane na zaprojektowaniu redukcji wymiarów w celu podniesienia gwarancji probabilistycznych oraz konstrukcji efektywnych algorytmów najblizszych sąsiadów dla różnych metryk.