Reg. No: 2018/29/N/ST6/00676; Principal Investigator: mgr Piotr Wygocki

Nearest neighbors problem without false negatives – synopsis for general public

Piotr Wygocki

June 5, 2018

1 Motivation

Assume that we have an extremely large collection of images. For a given query image, our aim is to determine the most similar image from the input set. One of the possible applications could be to use such queries for image classification. Assume that we want to assign each query image to one of the given categories. Our strategy might be to find the most similar image from already categorized input set and assign the query image to the same category.

The problem described above is called the nearest neighbor problem. In this algorithmic approach, we use the representation of each image as a tuple of hundreds of real numbers. The input data can be pre–preprocessed to build a data structure which will be used to answer queries.

One of the methods of finding the nearest neighbor is to scan the whole input set, compute the value of a given similarity measure and return the closest point. Unfortunately, this method is unacceptable for many practical purposes, where the query must be efficient. The problem can be solved when the dimension of our space is low. Unfortunately, under strong theoretical assumptions, the problem in high dimensional spaces cannot be solved efficiently, i.e., with the query time sub–linear in the number of input points and both the query time and the preprocessing time not exponential in the dimension of the space. Thus, the approximate near neighbor search was introduced.

The majority of the previously known algorithms satisfied Monte Carlo guarantees, i.e., an algorithm might return an incorrect answer with some probability.

2 Objective

Our main objective is to design an efficient algorithm with stronger guarantees, e.i., the Las Vegas guarantees. This means, that the found point is always the nearest neighbor. Moreover, existing algorithms often have many parameters which are hard to tune. In this project, we are planning to address this issue, which can impact various areas where the nearest neighbor techniques are applied, such as computer vision, robot sensing or pattern recognition.

3 Research

The research will focus on different areas of computational geometry. The nearest neighbor problem is well understood in low–dimensional case. Moreover, there are efficient algorithms even when the dimension of the input space is logarithmic in the number of input points. In view of the above, the natural idea is to reduce the dimension of the input space and solve the low–dimensional problem. Such dimension reductions are often implemented as linear projections. Unfortunately, these projections are usually probabilistic, i.e., each entry of the projection matrix is sampled independently from some probability distribution. Moreover, such projections preserve distances only with some probability (see e.g., well known Johnson Lindenstrauss Lemma). There are derandomization results but, up to the author's knowledge, none of them can be directly applied in the nearest neighbor setting. The research will be focused on understanding the random processes and designing processes which will enhance probabilistic guarantees.