Nearest Neighbor Search Data Structure

Sanjoy Dasgupta
Department of Computer Science
University of California, San Diego
dasgupta@cs.ucsd.edu

Zhen Zhai
Department of Computer Science
University of California, San Diego
zzhai@ucsd.edu

Eugene Che
Department of Computer Science
University of California, San Diego
eche@ucsd.edu

Abstract

Nearest-neighbor(NN) search is boardly used within all different fields of study to gain information on new data from training data set. For NN search, the more complex the training data set is the more accurate the result will be. However, doing NN search on complex and large training data sets is time consuming. Therefore, improving the speed and acuracy of nearest-neighbor search becomes essential. We look at different data structures for NN search and compare the results on varied data sets. We focus on data structures including kd tree, kd spill tree, kd virtual spill tree, bsp tree, and bsp spill tree. We conclude that spilling often improve the performance of the data structures.

1 Introduction

A lot of machine learning algorithms spend a big amount of time searching for information of an input query from the training set, which could very well be represented by high dimentional vectors. This approach is referred as nearest neighbor search. We are given a data set that contains a set of points

2 Data structures

We propose two main data structures, KD-trees and BSP-trees. We also look at KD spill tree, BSP spill tree, and KD virtual spill tree, which are the revised versions. We show that spilling improve the accuracy of NN search and virtual spill improve the performance even more.

2.1 KD-trees

2.2 BSP-trees

2.3 Spill trees and virtual spill trees

3 Experimental results

3.1 Mnist

3.2 CIFAR

3.3 eHarmony

4 Conclusion

5 References