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 and PCA tree. We also look at KD spill tree, KD virtual spill tree, and PCA spill tree to explore about the technique of spilling and virtual spilling. We conclude that spilling and virtual spilling improve the performance of the data structures.

1 Introduction

Gaining information of an input query by looking for the most similar match from the training set is referred as nearest neighbor(NN) search. We are given a data set **S** that contains a set of points $\{s_1, s_2, s_3, ..., s_n\}$ and an input query q, and we are looking for the nearest neighbor s_i of q in **S** such that s_i can provide valuable information about input query q.

A lot of machine learning algorithms spend a big amount of time on nearest neighbor search. The reason that NN search is time consuming is that the training set could very well be represented by high dimentional vectors. In fact, the accuracy of the NN search result is proportional with the complexity of the training set. Therefore, speeding up NN search can lead to tremoudous improvment to many applications of not only machine learning but also data analysis, bio-informatics, signal processing, etc.

There has been a large number of research devoting to speed up the performance of NN search. In this paper, we evaluate two most promising data structures, KD tree and PCA tree. (TALK ABOUT THESE TWO DATA STRUCTURE)

We also experiment the technique of spilling and virtual spilling by looking at KD spill tree, KD virtual spill tree, and PCA spill tree. (TALK ABOUT THESE TWO TECHNIQUE)

We apply these data structures on five different data sets from different area of study including computer vision, human relationship... We focus on the accuracy of NN search using these data structures and conclude the following: PCA tree has a better performance comparing to KD tree, spilling improve the performance, and virtual spilling improve the accuracy slightly more than spilling.

2 Data structures

We focus our study on binary space partitioning(BSP) trees, one of the most promising data structure for NN search. We first look at k-dimentional(KD) trees. Then we study principle component analysis(PCA) tree, in which we preprocess the data using PCA before partitioning the space. Finally, we look at spill trees and virtual spill trees, which are the revised data structures of BSP trees. Experiment supports that spilling improve the accuracy of NN search and virtual spilling improve the performance even more.

2.1 KD trees

KD tree is ...

Pseudocode:

2.2 PCA tree

PCA tree is ...

Pseudocode:

2.3 Spill trees and virtual spill trees

Spill tree is ...

Virtual spill tree is ...

3 Experimental results

3.1 Mnist

3.2 CIFAR

3.3 eHarmony

3.4 data4

3.5 data5

4 Conclusion

5 References