Nearest Neighbor Search in General Metric Spaces

Neel Rakholia,[1*] William March,[2] George Biros[2]

[1]Department of Applied Mathematics and Applied Physics,
Columbia University, New York, NY 10027, USA
[2]Institute for Computational Engineering and Science,
University of Texas at Austin, Austin, TX 78705, USA

[*]To whom correspondence should be addressed; E-mail: nvr2105@columbia.edu

Abstract. Significant work has been done on addressing the problem of nearest neighbor (NN) search in Euclidean Space. Notable is the wealth of literature on search techniques involving the use of trees. Kd-trees and Ball trees are among the more commonly used data structures to speed up NN search. Surprisingly little work however has been done on using trees to find nearest neighbors in general metric spaces.

In this paper we undertake a study of vantage point trees (VP-trees), and analyze their effectiveness in finding NN for kernel based distance metrics. We also propose two methods for searching these trees: an exact backtrack search algorithm, and an approximate random tree search algorithm. Previous work on VP-trees has focused on a priority queue based search, which is not very effective for even moderately high dimensional data. Our approach is unique in this regard.

For an RBF-kernel based distance metric, 500,000 training points and 54 features, NN search on 20,000 query points using the random trees search method yielded 90 percent accuracy with only about 1 percent of total distance evaluations: an improvement of about 2 orders of magnitude.

Keywords. Nearest Neighbor Algorithms, Tree Codes, Metric Spaces, Data Analysis, VP-Trees, Machine Learning

1 Introduction

The nearest neighbor problem refers to finding the set of points P_c in a database of points D that are closest to a query point q. The notion of close and correspondingly distance can be fairly arbitrary as long as it follows the following properties. For a space to be metric, these properties must hold true. [1]

1. Reflectivity: $d(a, a) = 0$
2. Symmetry: $d(a, b) = d(b, a)$

3. Non-Negativity: $d(a, b) > 0$, $a \neq b$

4. Triangle Inequality: $d(a, b) \leq d(a, c) + d(b, c)$

One of the more widely used distance metrics are derived from similarity measures such a kernels. These are of immense practical importance in fields such Computer Vision and Natural Language Processing where the concept of similarity between abstract objects is of importance. A kernel $K : \mathbb{R}^d \times \mathbb{R}^d$ is a similarity function with the property that for any x and y, the distance between x and y increases, $K(x, y)$ decreases. The construction of the kernel distance subsequently involves a transformation from similarities to distances. It can be represented in the following general form. Given two "objects" A and B, and a measure of similarity between them given by $K(A, B)$, then the induced distance between A and B can be defined as the difference between the self-similarities $K(A, A) + K(B, B)$ and the cross-similarity $K(A, B)$. Additionally this distance could be normalized by taking the square root. [2]

$$d(A, B) = \sqrt{K(A, A) + K(B, B) - 2K(A, B)} \tag{1}$$

VP-trees require the use of a bounded distance metric: a metric that yields distance between $[0, 1]$. Any unbounded kernel distance metric can be scaled to be a bounded metric by the following simple transformation: [3]

$$d'(A, B) = \frac{d(A, B)}{1 + d(A, B)} \tag{2}$$

Significance Vantage point trees (VP-trees) use concentric hyperspheres to partition points into a metric tree. [3]

References

[1] Neeraj Kumar, Li Zhang, and Shree Nayar. What is a good nearest neighbors algorithm for finding similar patches in images? In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, volume 5303 of *Lecture Notes in Computer Science*, pages 364–378. Springer Berlin Heidelberg, 2008.

[2] Jeff M. Phillips and Suresh Venkatasubramanian. A gentle introduction to the kernel distance. *CoRR*, abs/1103.1625, 2011.

[3] Peter N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1993.