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