Efficient Model Selection for Large-Scale Nearest-Neighbor Data Mining

Greg Hamerly and Greg Speegle

Baylor University, Waco, TX 76798, USA
greg_hamerly@baylor.edu, greg_speegle@baylor.edu

Abstract. One of the most widely used models for large-scale data mining is the k-nearest neighbor (k-nn) algorithm. It can be used for classification, regression, density estimation, and information retrieval. To use k-nn, a practitioner must first choose k, usually selecting the k with the minimal loss estimated by cross-validation. In this work, we begin with an existing but little-studied method that greatly accelerates the cross-validation process for selecting k from a range of user-provided possibilities. The result is that a much larger range of k values may be examined more quickly. Next, we extend this algorithm with an additional optimization to provide improved performance for locally linear regression problems. We also show how this method can be applied to automatically select the range of k values when the user has no *a priori* knowledge of appropriate bounds. Furthermore, we apply statistical methods to reduce the number of examples examined while still finding a likely best k, greatly improving performance for large data sets. Finally, we present both analytical and experimental results that demonstrate these benefits.

Key words: data mining, k nearest neighbor, optimal parameter selection